1994-06

Speech recognition of foreign accent

Dewey, John K.

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

SPEECH RECOGNITION OF FOREIGN ACCENT

by

John K. Dewey

June, 1994

Thesis Advisor:	Monique P. Fargues
Thesis Co-Advisor:	Ralph Hippenstiel

Approved for public release; distribution is unlimited

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE June 1994	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE Speech Recognition of Foreign Accent	5. FUNDING NUMBERS
6. AUTHOR(S) John K. Dewey	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited	12b. DISTRIBUTION CODE A

13. ABSTRACT *(maximum 200 words)*

This thesis investigates the application of AutoRegressive (AR) modeling techniques on single syllable words to detect foreign accents in spoken American English. The study involves thirty-one native American English speakers, and six native Brazilian speakers. Five different distance measures are used for classification. Results show that correct classification is obtained for 88 % of the native English speakers and 80.5 % of the non-native (foreign) English speakers.

14. SUBJECT TERMS Speech Processing, Foreign Accent Recognition, AutoRegressive (AR) modeling	15. NUMBER OF PAGES 83
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std 239-18

Approved for public release; distribution is unlimited.

Speech Recognition of Foreign Accent

by

John K. Dewey
Captain, United States Army
B.S., Wright State University, 1985

Submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
June 1994

Author: [redacted]
John K. Dewey

Approved by: [redacted]
Monique P. Fargues, Thesis Advisor

[redacted]
Ralph Hippenstiel, Thesis Co-Advisor

[redacted]
Michael A. Morgan, Chairman
Department of Electrical and Computer Engineering

ABSTRACT

This thesis investigates the application of AutoRegressive (AR) modeling techniques on single syllable words to detect foreign accents in spoken American English. The study involves thirty-one native American English speakers, and six native Brazilian speakers. Five different distance measures are used for classification. Results show that correct classification is obtained for 88 % of the native English speakers and 80.5 % of the non-native (foreign) English speakers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

The completion of this thesis and my master's of electrical engineering degree would not have been possible without my consummate friend, untiring supporter, and wife Mary. My wholehearted appreciation goes out to our children, Elise and John, my parents, Willard W. and Sondra Dewey, and my parents in law, Patrick and Edith Causey whose love and support continuously motivated me. Additionally, I must thank my thesis advisors, Professor Monique Fargues, Professor Ralph Hippenstiel, and the Naval Postgraduate School Faculty for their devotion and teachings.

I. INTRODUCTION

The goal of accent recognition investigated in this thesis is to automatically detect non-native (foreign) English speakers as foreign, and native American English speakers as native. Automatic recognition refers to the ability to detect foreign accents using computers or machines. The detection of foreign accents by ear is common practice. However, the automatic detection of foreign accents is difficult due to the time varying frequencies in normal speech and the potential bias of loudness, and how fast individuals speak.

This thesis considers the use of a few single syllable words common in daily speech. A normalization technique limits the effects of loudness, and how fast individuals speak. This study focuses on one group of non-native English speakers with the notion that the techniques used for accent detection may be extended to recognize non-native English speakers from many languages. The group selected for this study consists of Brazilian students attending the Naval Postgraduate School. The word list used is made up of words that are difficult for native Brazilians to pronounce. This word list selection process is based on the idea that "You Can't Teach Old Dogs New Tricks" [1, 2] and that the sounds used in native American English that are different from those sounds used by native Brazilians will be more often mispronounced. The native English speakers used in this study are originally from various regions of the United States and are all military

servicemen which limits regional accent due to the many areas of their travels and residences. The techniques described in this study may enhance the ability to recognize foreign accents and enable language schools to test student accents automatically. Additionally, the ability to recognize foreign accents has broad military use.

The remainder of this thesis is organized as follows: Chapter II introduces speech analysis and presents a brief introduction to phonetic concepts. Chapter III introduces AutoRegressive (AR) modeling. Chapter IV presents the method of data collection, preparation, and normalization used for this study. Chapter V presents the premise of foreign accents and word list selection. Chapter VI introduces the performance measures used to test the various speakers. Results are presented in Chapter VII. Finally, Chapter VIII presents conclusions and recommendations for future research.

II. SPEECH ANALYSIS

This chapter first explains how speech may be divided into individual sounds and combinations of sounds. Next, speech terminology is introduced with brief explanations and definitions [3]. Finally, speech analysis techniques used to obtain information from speech signals are presented.

Speech signals have a special quality that most other signals do not have; their contents are usually recognizable to the listener even if the listener does not know what to expect. In addition, the quality or noisiness of the signal is usually immediately apparent, while the quality of other signals, such as tones or groups of tones, would not be as apparent to the lay listener. Speech is made up of many sounds created by many different mechanisms of articulation. This means that although every person sounds a little different, and even though there are many accents in normal speech, the various speech signals are still understandable among speakers speaking the same language. Linguistics is the scientific study of language and the manner in which these rules are used in human communication. The study of the abstract units and their relationships in a language is called phonemics, while the study of the actual sounds of the language is called phonetics. Phonemes are the basic theoretical unit for describing how speech conveys linguistic meaning (for example: the word "man" is constituted of three phonemes /m/, /æ/, /n/). The English language has forty-two phonemes which are listed in Table 1 [4]. English, in this study, refers to American English. Phonemes are defined as theoretical or ideal

TABLE 1 AMERICAN ENGLISH PHONEMES [Carrell, J., and Tiffany W., *Phonetics: Theory and Application to Speech Improvement*, McGraw-Hill 1960, reproduced with permission from the Publisher]

Vowels

Front vowels				Back vowels			
SYMBOL	KEY			SYMBOL	KEY		
[i]	heed	[hid]		[u]	who'd	[hud]	
[ɪ]	hid	[hɪd]		[ʊ]	hood	[hʊd]	
[e]	hayed	[hed]		[o]	hoed	[hod]	
[ɛ]	head	[hɛd]		[ɔ]	hawed	[hɔd]	
[æ]	had	[hæd]		[ɑ]	hod	[hɑd]	
Central vowels				**Diphthongs†**			
[ɝ-ɹ]*	hurt	[hɝt]		[aɪ]	file	[faɪl]	
[ʌ]	hut	[hʌt]		[aʊ]	fowl	[faʊl]	
[ɚ-ə]*	under	[ʌndɚ]		[ɔɪ]	foil	[fɔɪl]	
[ə]	about	[əbaʊt]		[ju]	fuel	[fjul]	
Consonants							
Stops				**Fricatives**			
[p]	pen	[pɛn]		[f]	few	[fju]	
[b]	Ben	[bɛn]		[v]	view	[vju]	
[t]	ten	[tɛn]		[θ]	thigh	[θaɪ]	
[d]	den	[dɛn]		[ð]	thy	[ðaɪ]	
[k]	Kay	[ke]		[h]	hay	[he]	
[g]	gay	[ge]		[s]	say	[se]	
[tʃ]	chew	[tʃu]		[ʃ]	shay	[ʃe]	
[dʒ]	Jew	[dʒu]		[z]	bays	[bez]	
				[ʒ]	beige	[beʒ]	
Nasals and lateral				**Glides**			
[m]	some	[sʌm]		[w]	way	[we]	
[n]	sun	[sʌn]		[hw]	whey	[hwe]	
[ŋ]	sung	[sʌŋ]		[j]	yea	[je]	
[l]	lay	[le]		[r]	ray	[re]	

sounds, and if every speaker produced these ideal phonemes, English speech would be a simple combination of the phonemes. Phones are defined as the actual sounds produced by speakers which lead to the understanding of the intended meaning of the sounds. A phoneme spoken individually is simple to identify, however when phonemes are spoken in normal speech, the beginnings and ends of phonemes are very difficult to identify. In addition phoneme sounds may interact with each other. In normal speech, there are transition periods between phonemes where slight acoustic variations occur. Therefore, with each phoneme is associated a group of these transitional phone variations called allophones.

The basic phonemes in speech are made up of vowels (front, back, central), semivowels, diphthongs, fricatives, affricates, stops, glides, and nasals. Speech is also classified as voiced and unvoiced. Voiced and unvoiced speech can be separated using a combination of two speech analysis techniques called; zero crossing measure, and short-term energy measure [3]. The zero crossing measure identifies the number of times a sequence changes signs, and the short-term energy measure is used to determine where the sequences majority of the energy is located. Unvoiced speech are usually high frequency sounds that have large numbers of zero crossings and voiced speech normally contain the majority of the energy. Since the zero crossing measure, and short-term energy measure identify these characteristics, when used in combination, separation of the voiced and unvoiced speech is possible. Voiced speech are sounds that are created with a vocal note or sonat (as in the vowel sound in "sat" phonetically spelled [sæt]). Vowel

sounds are quasi-periodic and this period is known as the pitch period. Unvoiced speech are sounds that are whispered or created without vocal note (as in the constants sounds in "sat"). The classes of vowels get their names from how they are articulated, or how the tongue is used to produce a sound. Semivowels are vowel-like sounds not caused by vowels (the m and n sounds in "man"). Diphthongs are sometimes called long vowels, however they are actually the sounds created when transitioning from one vowel sound to another in a continuous fashion (as in "being" or "seeing"). Fricatives are voiced or unvoiced noise-like sounds used in speech (for example: /z/ and /v/ are voiced phonemes while /s/ and /f/ are unvoiced phonemes). Stops or plosives are constant sounds that are normally aspirated in English, and where a release of air under pressure accompanies the sound (for example: b, d, g, p, t, k). Affricates are formed by transitioning from a stop to a fricative (as in "church" and "John"). Continuant sounds like vowel sounds are quasi-periodic, and their frequency components can be captured using techniques that rely on stationarity. Time-varying sounds like those found in diphthongs and semivowels are non-stationary and are classified noncontinuant.

Figure 1 shows the recorded speech signal "being". Figure 2 shows the frequency spectrum of the same signal ("being"). The time-frequency spectrogram of "being" is shown in Figure 3, where the time increment is 3.4 milliseconds, and the Fast Fourier Transform (FFT) length is 512. The time varying voiced components of the frequencies are obvious in Figure 3. Thus, the spectrogram shows that there are definite advantages to looking at both the frequency spectrum and the spectrogram of a signal. Both the

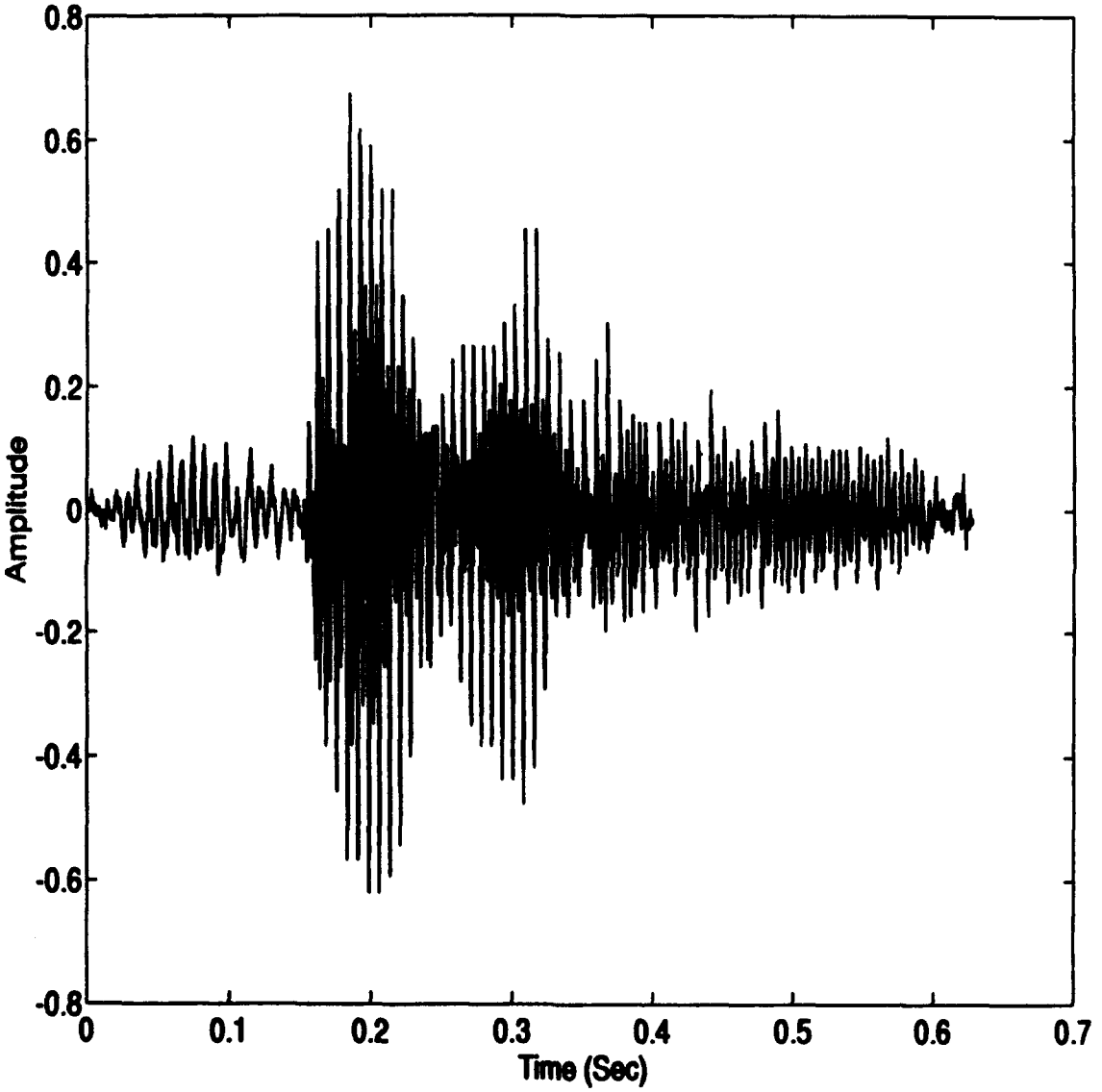


Figure 1: Recorded speech signal "being", sampling frequency $fs = 8192$ Hz.

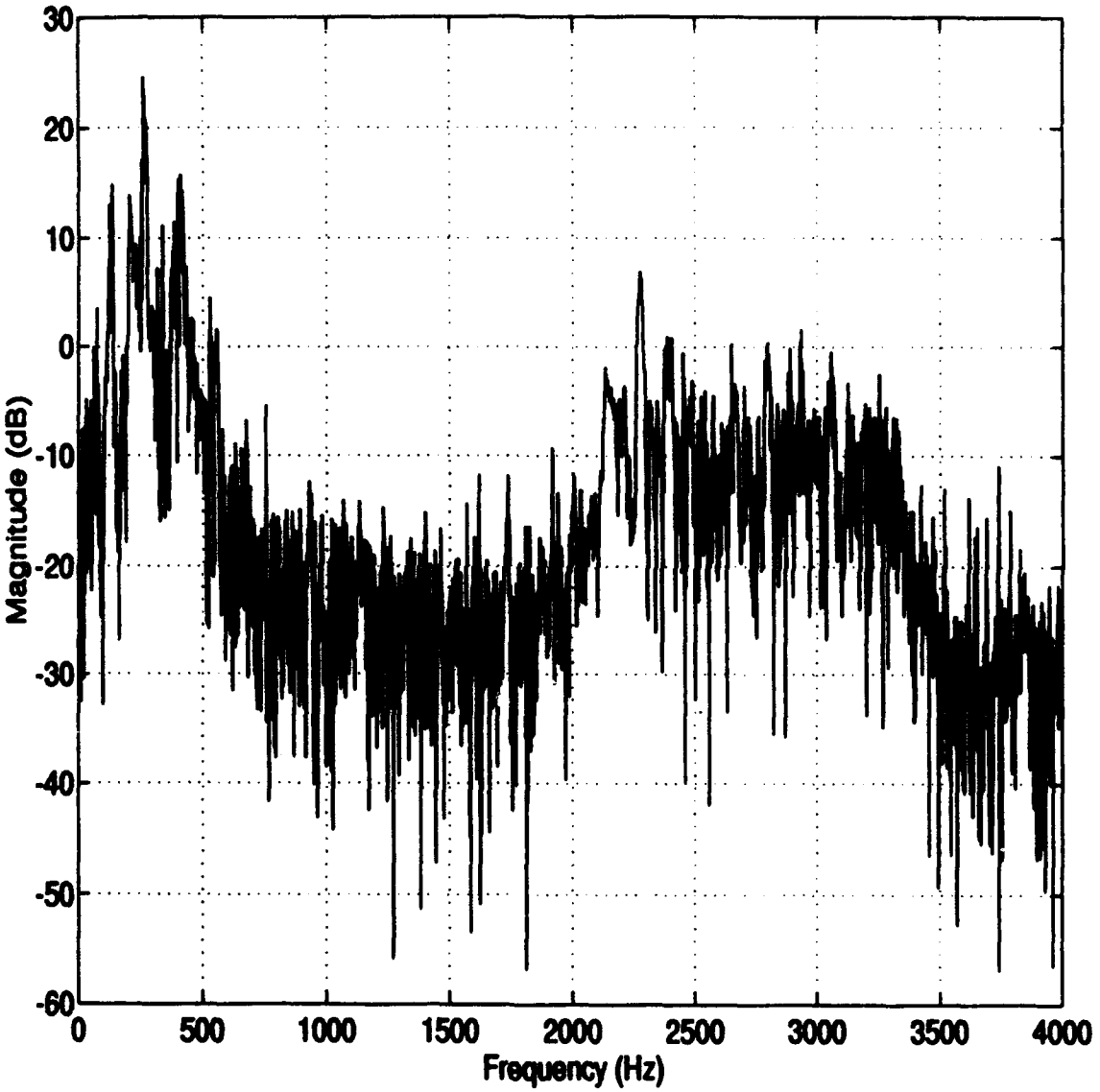


Figure 2: Frequency spectrum of the recorded speech signal "being".

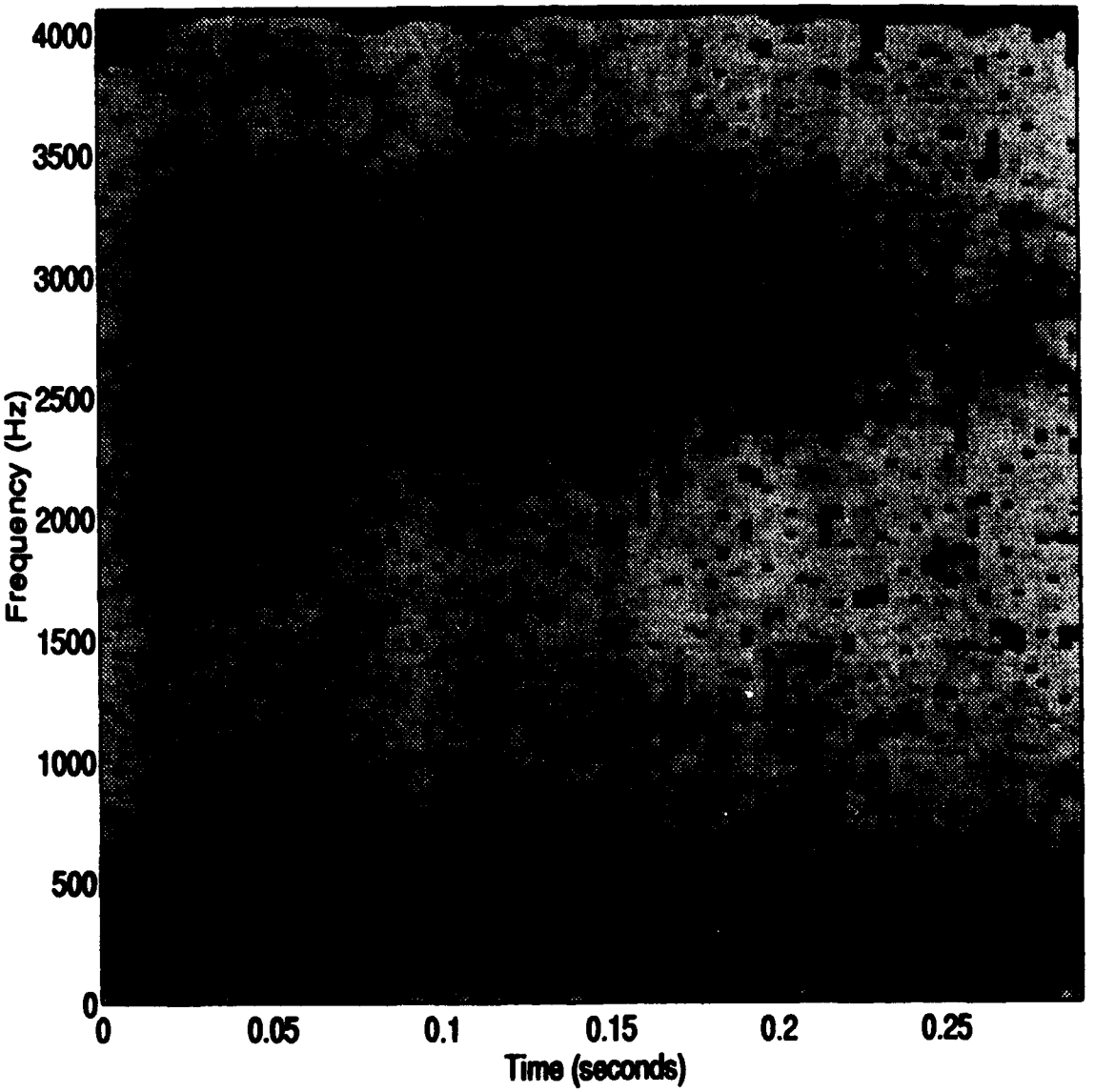


Figure 3: Time-frequency spectrogram of the recorded speech signal "being", where the time increment is 3.4 ms., and the Fast Fourier Transform (FFT) length is 512.

strong center frequency components of the vowel-like sounds are identified, as well as the variation in the frequency over time which causes different sounds.

Note that English is a spoken language as opposed to a written language, meaning that groups of letters are not always pronounced the same way. As a result, symbols are used to express phoneme sounds. These symbols used to represent speech sounds are called phonetic symbols, and with these symbols the English language may be represented as a written language. There are several phonetic alphabets used for English pronunciation. Table 2 shows five common alphabets [4].

TABLE 2 FIVE COMMON AMERICAN ENGLISH PHONETIC ALPHABETS

IPA	Webster's New Collegiate	Webster's New World	American College	NBC Handbook	IPA	Webster's New Collegiate	Webster's New World	American College	NBC Handbook
i	ē	ē	ē	ee	k	k	k	k	k
ɪ	ĭ	i	ĭ	i	g	g	g	g	g
e	ā	ā	ā	ay	tʃ	ch	ch	ch	ch
ɛ	ĕ	e	ĕ	ai e	dʒ	j	j	j	j
æ	ă	a	ă	a	f	f	f	f	f
ɑ	ŏ	ä o	ä	ah	v	v	v	v	v
ɔ	ô	ô	ô	aw	θ	th	th	th	th
o	ō	ō	ō	oh	ð	th	th	th	th:
U	o͝o	oo	o͝o	oo	s	s	s	s	s
u	o͞o	o͞o	o͞o	oo:	z	z	z	z	z
ʌ	ŭ	u	ŭ	uh	ʃ	sh	sh	sh	sh
ɜ	ûr	ŭr	ûr	er	ʒ	zh	zh	zh	zh
ə	(italics)	ə	ə	uh	h	h	h	h	h
ə	ĕr	ĕr	ər	er	m	m	m	m	m
aɪ	ī	ī	ī	igh	n	n	n	n	n
ɔɪ	oi	oi	oi	oi	ŋ	ng	ŋ	ng	ng
ju	ū	ū	ū	yoo:	l	l	l	l	l
aU	ou	ou	ou	ow	w	w	w	w	w
p	p	p	p	p	hw	hw	hw	hw	hw
t	t	t	t	t	j	y	y	y	y
b	b	b	b	b	r	r	r	r	r
d	d	d	d	d					

Most dictionaries include a phonetic pronunciation with each word. The phonetic alphabet used in this study is the International Phonetic Alphabet (IPA).

The voiced speech phonemes are quasi-periodic, in Figure 4 (a blown-up section of Figure 1) the quasi-periodic nature is shown. The fundamental period T = 7.5 ms, of the waveform shown in Figure 4, is called the pitch period. The nominal center frequencies of the resonances present in the voiced speech phonemes are called formant frequencies, or formants. These frequencies would be considered normal speech, or in this case theoretical or ideal frequencies. The first three formant frequencies for a voiced phoneme are normally labeled F1, F2, and F3. Table 3 shows some basic voiced phonemes and their associated average adult male formant frequencies [5].

TABLE 3 AVERAGE MALE FORMANT FREQUENCIES

Phonemes	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʊ/	/u/	/ʌ/	/ɝ/
Formants (Hz)										
F1	270	390	530	660	730	570	440	300	640	490
F2	2290	1990	1840	1720	1090	840	1020	870	1190	1350
F3	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690

Refer to Table 1 for examples of the sounds listed in Table 3. Note that male and female formant frequencies are very different on average. Therefore to eliminate problems due to gender differences, this study uses only adult male voices. In addition, age may also create some frequency discrepancies in formants, especially between children and adults. The

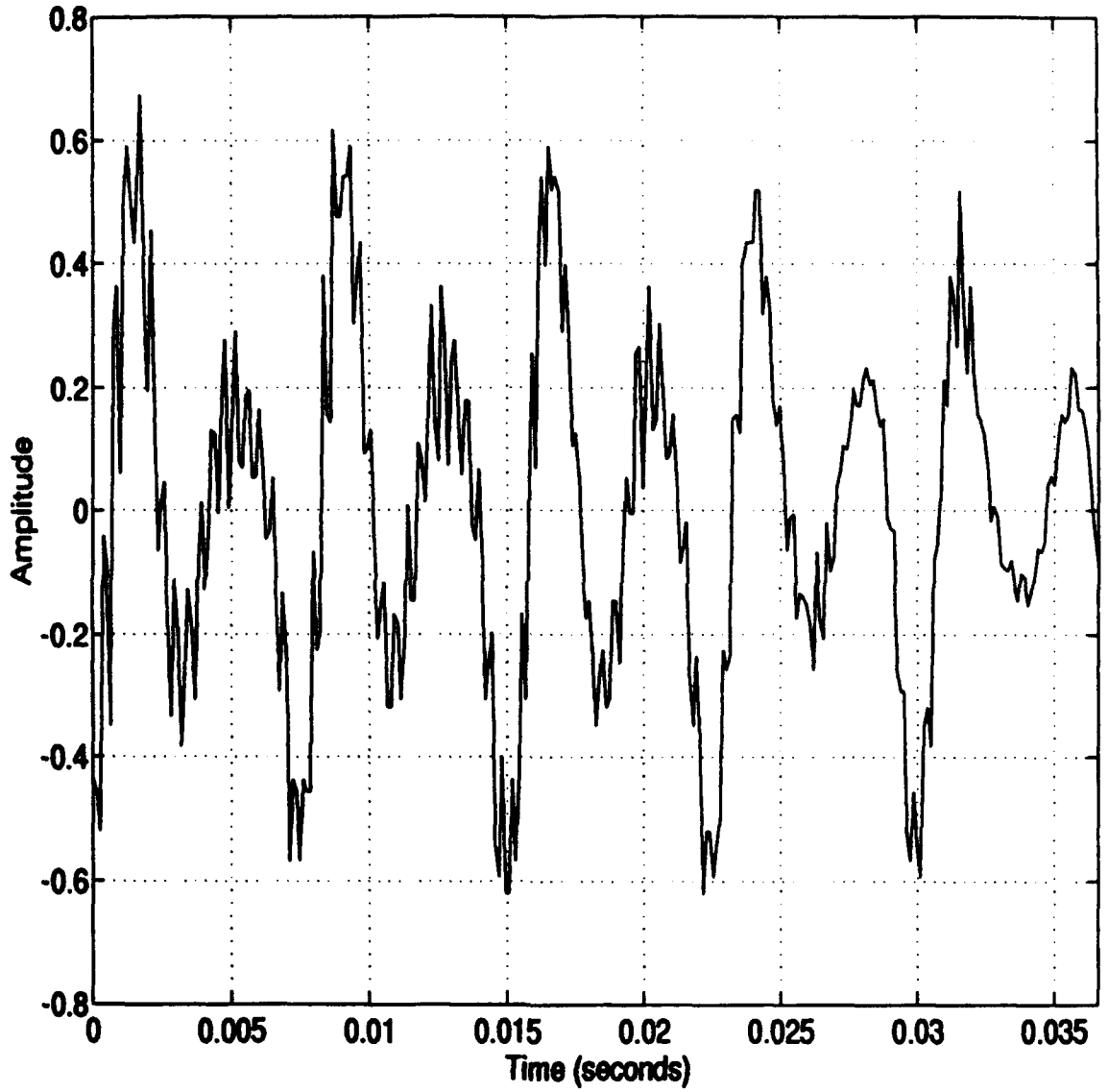


Figure 4: Blown-up section of Figure 1 showing the quasi-periodic nature of voiced speech phonemes. The pitch period is $T = 7.5$ ms.

database generated for this study uses only males ranging in age between twenty-eight and forty with an average age of thirty-one. The table of formants (Table 3) by no means is inclusive and does not begin to represent the phones or allophones. Figure 5 shows the frequency spectrum of the phoneme /æ/ produced by a native English speaking male. Recall that the first three ideal average formant frequencies for the phoneme /æ/ are located at 660 Hz, 1720 Hz, and 2410 Hz, and are indicated on Figure 5. Note that this speaker's second and third formant frequencies F2 and F3 are higher than the representative F2 and F3 averages.

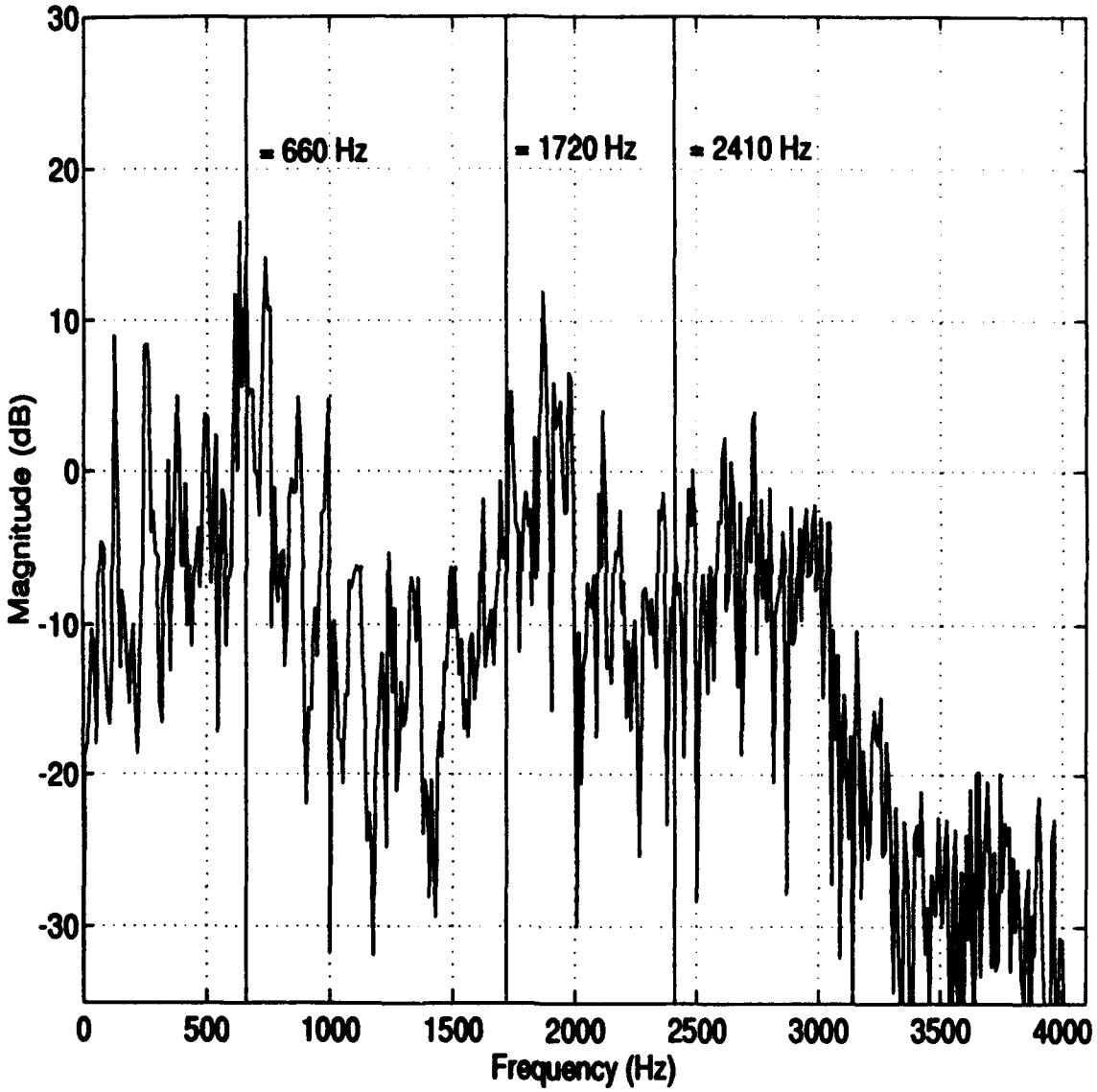
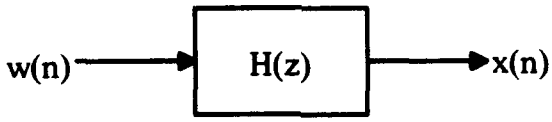


Figure 5: Frequency spectrum of the phoneme /æ/ produced by a native English speaking male.

III. AUTOREGRESSIVE MODELS

System modeling has been investigated extensively in signal processing applications. Models can be used for various applications such as; prediction, forecasting, and data compression. One of the most used linear models is the AutoRegressive (AR) model [6]. In the AR model a signal $x(n)$ is considered to be the output of some system with input $w(n)$ where $w(n)$ is white noise with power $\sigma_w^2 = 1$. The block diagram of the system is given by:



and the difference equation is given by:

$$x(n) = -a_1 x(n-1) - a_2 x(n-2) - \cdots - a_P x(n-P) + b_o w(n). \quad (1)$$

The coefficients, a_k for $k = 1, \cdots, P$, and b_o, are the parameters of the system, and P is the order of the AR model. The frequency domain expression obtained from the Z transform of (1) is given by:

$$X(z) = -a_1 z^{-1} X(z) - a_2 z^{-2} X(z) - \cdots - a_P z^{-P} X(z) + b_o W(z). \quad (2)$$

collecting like terms in (2) leads to:

$$X(z)[1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_P z^{-P}] = b_o W(z).$$

Let us define the polynomial:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_P z^{-P}.$$

For this study $w(n)$ is white noise with variance equal to one, the resulting transfer

function of the AR model is given by:

$$H(z) = \frac{X(z)}{W(z)} = \frac{b_o}{A(z)}.$$

The AR coefficients can be obtained by solving a set of linear equations obtained from equation (1). Using the properties of the AR model, the correlation function $R_x(l)$ obtained from $x(n)$ is given by:

$$R_x(l) = -a_1 R_x(l-1) - \cdots - a_P R_x(l-P) + b_o R_{wx}(l) ,$$

which leads to:

$$R_x(l) + a_1 R_x(l-1) + \cdots + a_P R_x(l-P) = b_o R_{wx}(l). \tag{3}$$

The cross correlation $R_{xw}(l)$ can be expressed in terms of the impulse response $h(n)$ of the AR system:

$$R_{xw}(l) = h(l) * R_w(l). \tag{4}$$

Recall that the correlation function of white noise is expressed as:

$$R_w(l) = \sigma_w^2 \delta(l). \tag{5}$$

Thus, substituting (5) into equation (4) leads to:

$$R_{xw}(l) = h(l) * \sigma_w^2 \delta(l) = \sigma_w^2 h(l),$$

which leads to:

$$R_{wx}(l) = \sigma_w^2 h^*(-l). \tag{6}$$

Next, substituting equation (6) into equation (3) leads to:

$$R_x(l) + a_1 R_x(l-1) + \cdots + a_P R_x(l-P) = b_o \sigma_w^2 h^*(-l) .$$

$h(n)$ is the impulse response of a causal filter, where a causal system produces output values which are expressed in terms of past and present input values only. Thus, $h(n)$ for $n < 0$ is equal to 0. Next using the Initial Value Theorem, we have:

$$h(0) = \lim_{z \to \infty} H(z) = \lim_{z \to \infty} \frac{b_o}{1 + a_1 z^{-1} + \cdots + a_P z^{-P}} = b_o$$

therefore, $R_{wx}(l) = b_o * \sigma_w^{\,2}$ for $l = 0$
$R_{wx}(l) = 0$ for $l > 0$

Expressing (3) for $l = 0, \cdots, P$ leads to the following system of linear equations known as the Yule-Walker equations:

$$\begin{bmatrix} R_x(0) & R_x(-1) & \cdots & R_x(-P) \\ R_x(1) & R_x(0) & \cdots & R_x(-P+1) \\ \vdots & \vdots & \vdots & \vdots \\ R_x(P) & R_x(P-1) & \cdots & R_x(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \sigma_w^2 |b_o|^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} |b_o|^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Figure 6 shows the spectral response obtained from an AR model of order twenty-four $P = 24$ for the speech signal "being", superimposed on the FFT spectrum of the same signal. The spectrum of an AR model is the magnitude of the frequency response of the AR model's transfer function. Note that the AR model more closely approximates the portion of the spectrum with high energy content, which are due to "pole-like" behavior, than it approximates the portion of the spectrum where the energy is lower.

The vowel sounds contained in words are quasi-periodic voiced components. The vowel frequencies contain the majority of the power in a single spoken word. The assumption here is that for single short words the most distinguishable components would then be the vowel-like sounds, and therefore the overall AR model of a word is a "good" representation. However, note that the AR model represents in some sense the "average" frequency information contained in the word, the non-stationary information present in the word cannot be represented by constructing the AR model of a full word. For example, results show that differences in the resulting AR models of single syllable full words and the voiced phonemes present in those words are very small. Figure 7 shows the closeness of the AR models for the phoneme /æ/, and the full word "sat", which contains the phoneme /æ/ where an AR model of order twenty-four is used. Full word AR

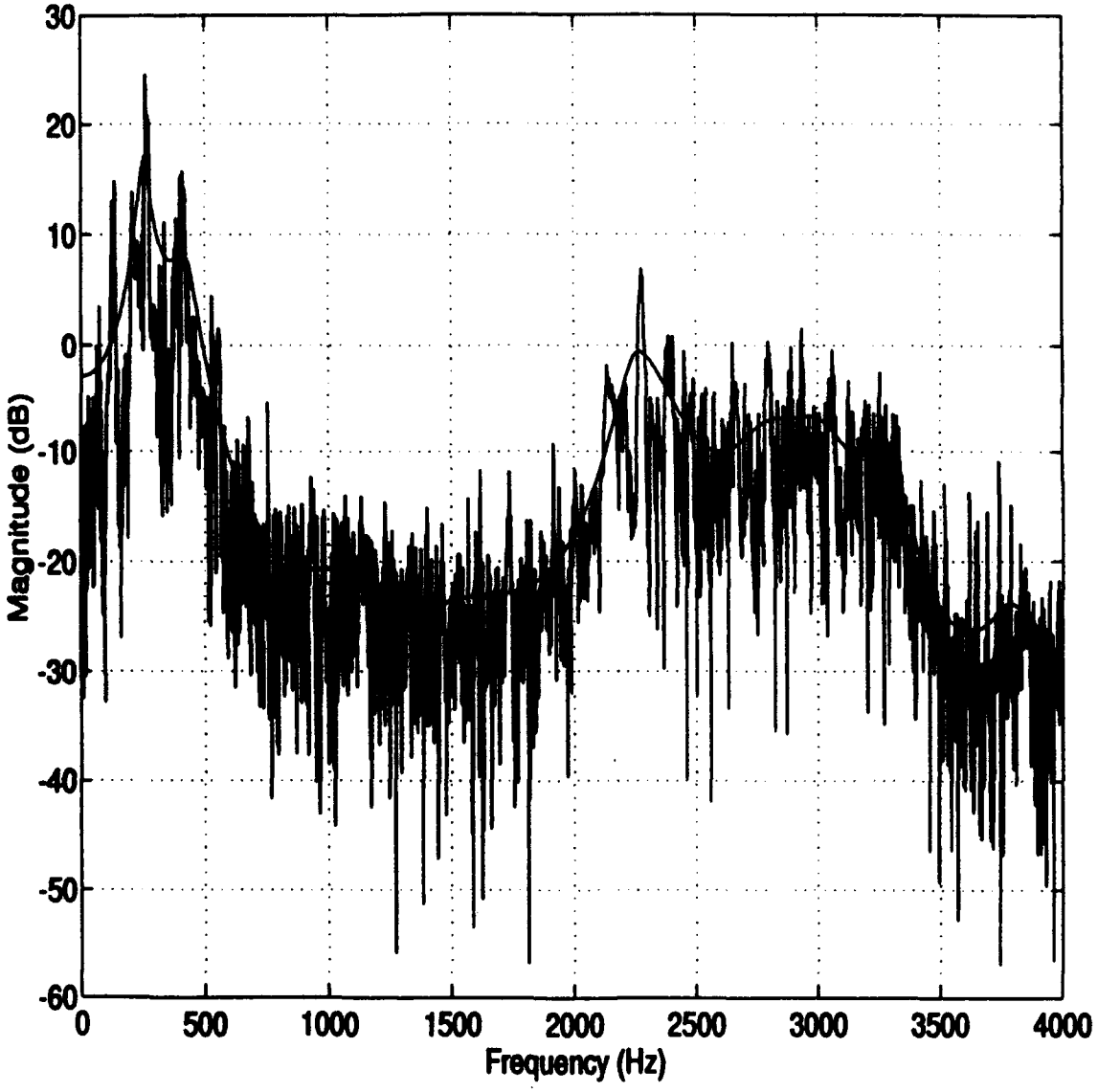


Figure 6: AR (smooth line) and FFT (jagged line) spectra of the recorded speech signal "being", the correlation method is used to compute the AR model, AR model order is $P = 24$.

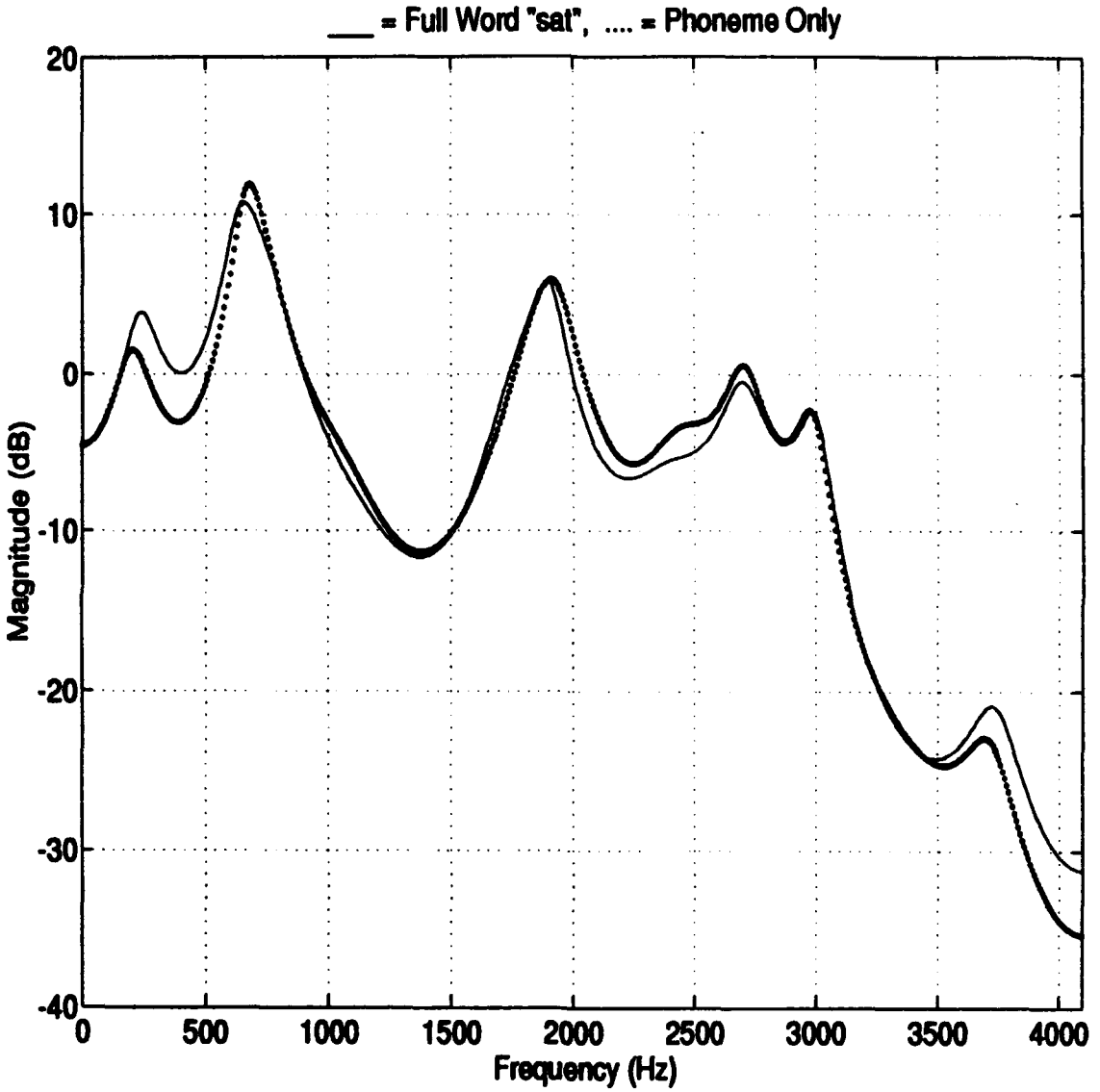


Figure 7: AR spectra for the phoneme /æ/, and the full word "sat", AR models are computed using the correlation method, AR models order are $P = 24$.

models are much easier to produce since no word segmentation is required. For the purpose of this study full word AR models are used. The AR model can be used to locate the first three formants F1, F2, and F3. The formants however, are not equally energy weighted. The lower frequencies usually contain more energy than the higher frequencies, and therefore only the frequency range where the first two formants are usually located is considered in this study. Table 3 shows that for any vowel the highest frequency for the second formant is around 2290 Hz, associated to the phoneme /i/. As a result, this study is restricted to the frequency range from 0 to 2400 Hz to consider the effects due to the first two formants only.

The order of the AR model was determined heuristically through experimentation. Table 3 is used, and AR models representing the words containing the phonemes of interest are produced to express the formant frequencies. An order of twenty-four is high enough to represent the spectral information contained. This order may appear to be large, however it allows a representation of enough details, while a lower order model may cause more information to be lost. Figure 8 shows the twelfth order AR model of the word "girl", and Figure 9 shows the twenty-fourth order AR model of the same sequence. Comparing the models obtained for order twelve and twenty-four in Figures 8 and 9 show that more details are represented with the higher order model.

The MATLAB™ implementation of the AR spectra is presented in Appendix A.

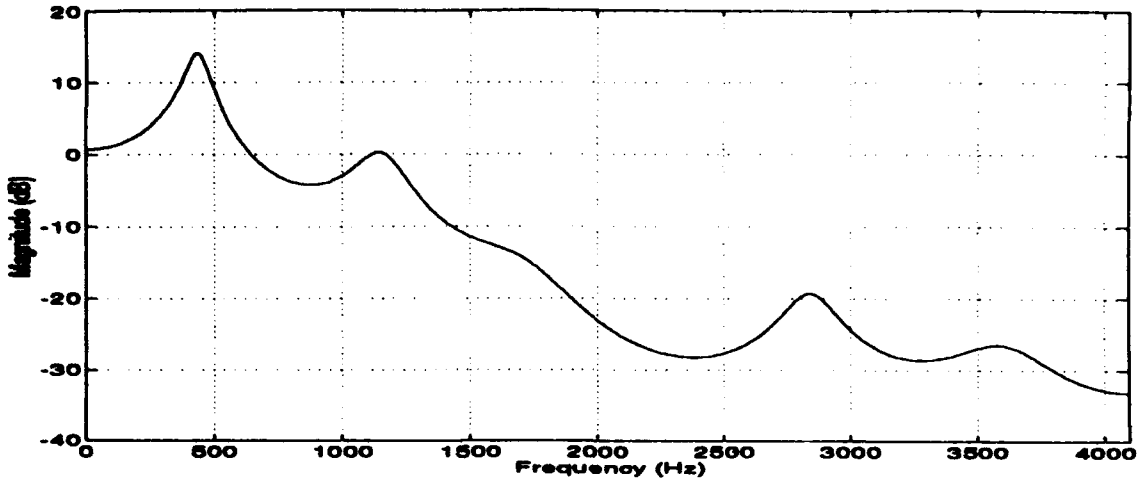


Figure 8: 12th order AR spectrum of the recorded speech signal "girl", correlation method used to compute the AR model.

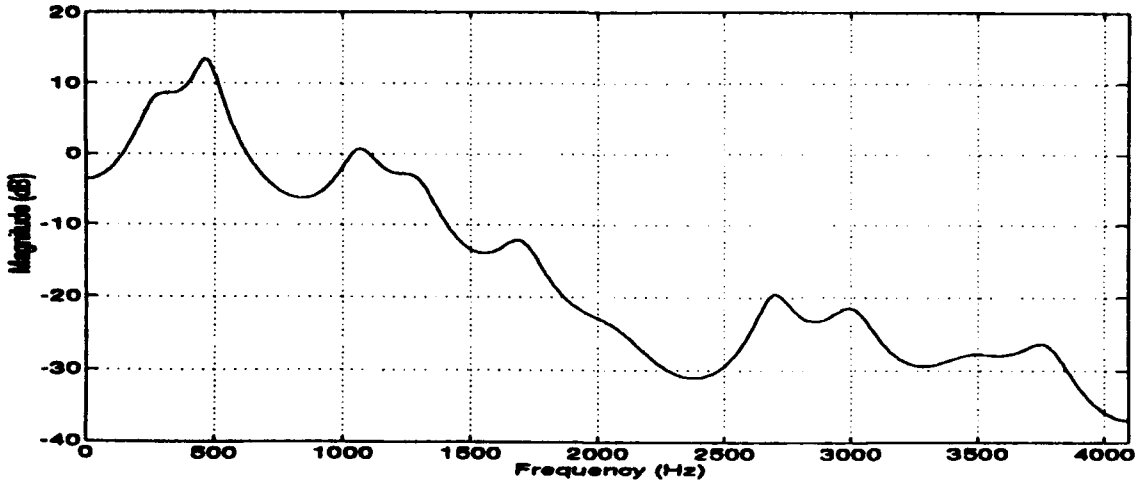


Figure 9: 24th order AR spectrum of the recorded speech signal "girl", correlation method used to compute the AR model.

IV. DATA PREPARATION

A group of thirty-one male native English speakers are recorded to represent a native English speaking model and a native English speaking test group. The thirty-one English speakers are divided into two groups; a model group of sixteen speakers, and a test group of fifteen speakers. Thirty-four English speakers were initially recorded. However, three were eliminated due to over and under-modulation or an unexplained anomaly in recordings. The resulting set of thirty-one native English speakers has an average age of thirty-one.

A second group is formed of non-native English speakers consisting of six male Brazilians with a native language of Portuguese. All Brazilian non-native English speakers are students at the Naval Postgraduate School pursuing graduate degrees. The Brazilian group has an average age of thirty-two, and on average all the individuals in that group have spoken English for more than thirteen years.

The software package used for numeric computation and graphics is MATLAB™.

A. RECORDINGS

All native and non-native English speakers are recorded in the same way. A Sun Sparc-10 workstation with an audio tool is used to directly record a list of fourteen spoken English words. Each speaker is recorded in the same room using identical equipment. The instructions given to the speakers are to relax, speak using their normal

voice, and pause momentarily between words to make word segmentation easier. The word list is reviewed by each speaker before the recording is started to ensure that every word on the list is understood. Each speaker is recorded saying the list of words twice. After the word list is recorded for the first time, the data file is saved and the process repeated. The word lists are digitized as recorded using a sampling frequency of 8192 Hz.

B. WORD SEPARATION

The process of data preparation begins with loading each data sequence, a list of fourteen spoken words, into MATLAB™. The word list is then plotted and cut into individual words visually. Each word is saved as a separate data file, and excess non-speech is trimmed from each spoken word. When cutting and trimming is completed, the separated list of fourteen words is saved as a data set. The resulting word data sequences consist of single spoken words with little excess silence before or after the word. Each speaker contributes two complete sets of data from the two times the word list is recorded.

C. FILTERING AND NORMALIZING

Each word data file is filtered and normalized before any processing begins.

1. Filtering

The normal speech frequency range is between 100 and 4000 Hz. A high-pass Finite Impulse Response (FIR) filter [7] with a cutoff frequency equal to 100 Hz is designed to eliminate the sixty Hertz equipment noise. Figure 10 shows the frequency

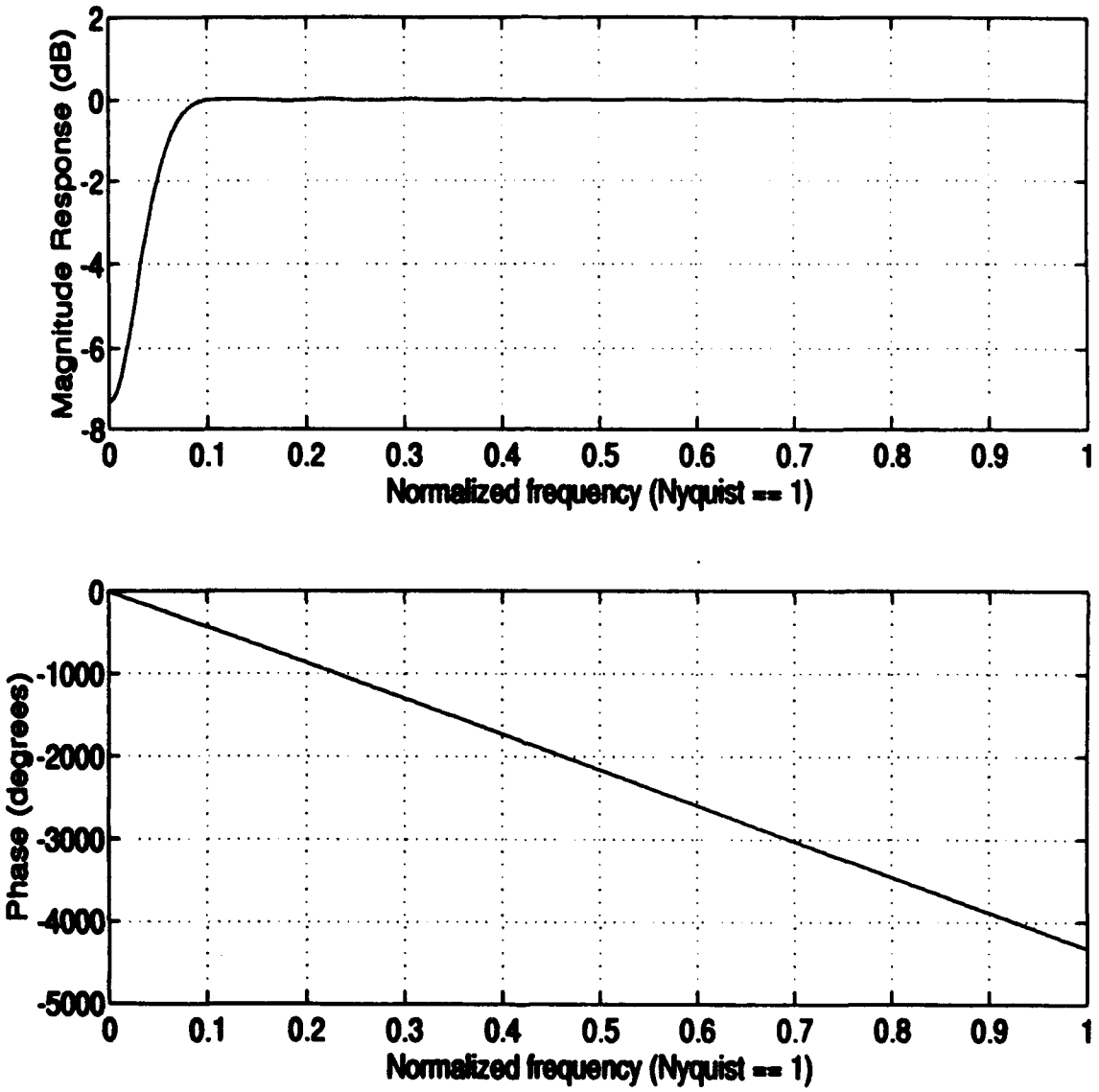


Figure 10: High pass, 48th order, FIR (Finite Impulse Response) filter with pass band frequency equal to 100 Hz. The sampling freuqncy is $fs = 8192$ Hz.

response of the high-pass filter. A low-pass Butterworth filter [7] is designed to eliminate all the frequencies above 4000 Hz. Figure 11 shows the frequency response of the low-pass filter.

2. Normalization

To achieve a goal of standard comparison between the spoken words, an energy normalization is required. Note that each word is spoken at different loudness levels and over different periods of time, as some speakers speak faster than others. Even though the environment of the recordings is constant, the speakers are very different. Each word data sequence has its mean removed before, between filters, and after the filtering is complete. To minimize the effects of loudness and variations in time or sequence length of the recorded speech, the following normalization is used:

$$\underline{Ndata} = \frac{\underline{Fdata}}{\sqrt{\underline{Fdata} \cdot \underline{Fdata}^T}},$$

where $\underline{Ndata}$ is the normalized word data sequence, and $\underline{Fdata}$ is the filtered word data sequence.

As a result, all AR models of a normalized word data sequence have the same energy without regard to speaker or word spoken. A check of the normalization can be conducted by finding the energy in the normalized word data sequence. The energy in each word data sequence is equal to one.

The result of the data preparation is a data sequence that can be comparatively analyzed with other data sequences prepared in the same manner. The potential effects of

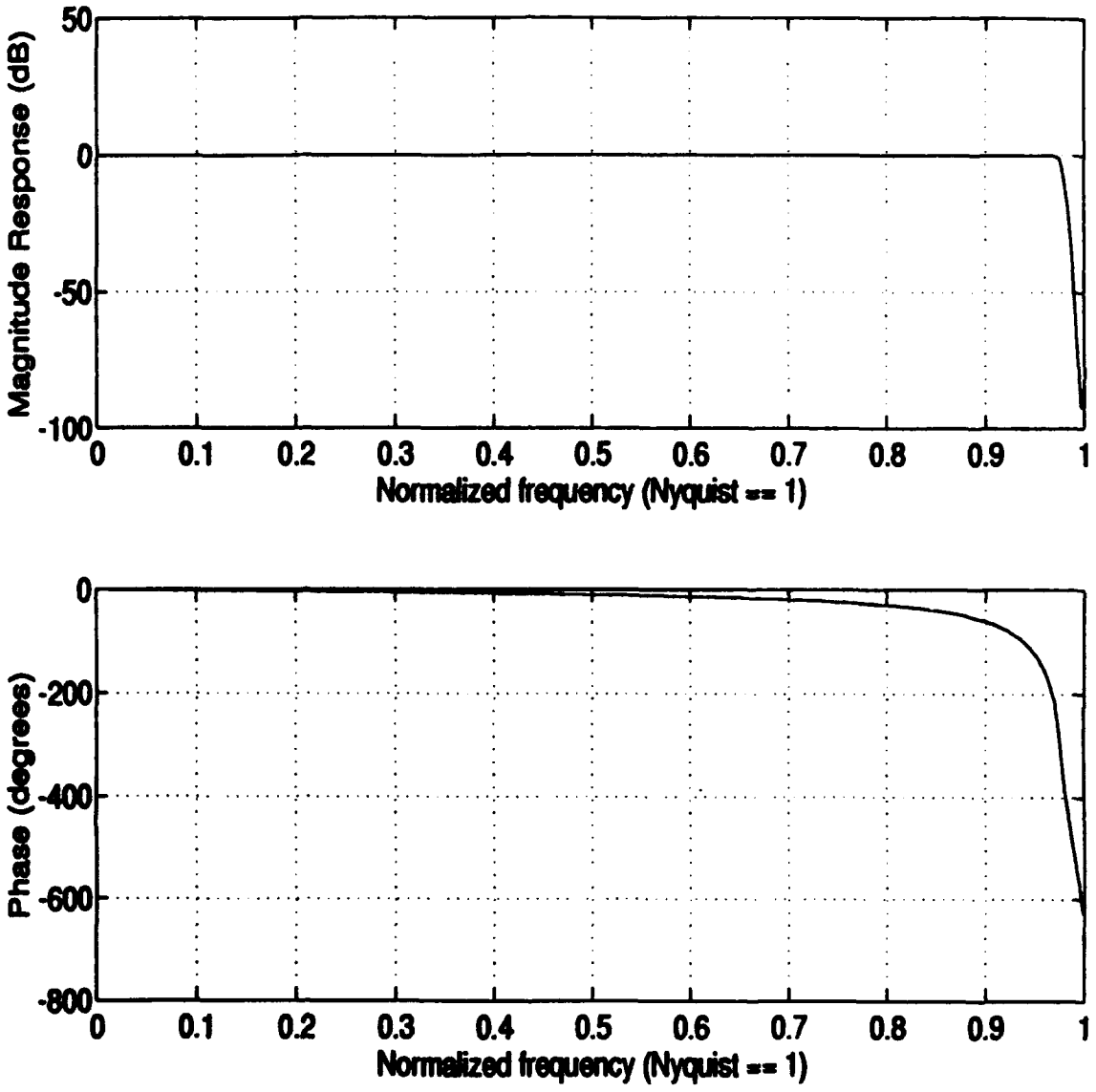


Figure 11: Low pass, 8th order, Butterworth filter with cut-off frequency equal to 4000 Hz. The sampling frequency is $fs = 8192$ Hz.

loudness variations and data sequence length of speech have been reduced, and the effects of energy variations and frequency concentration have been enhanced.

V. FOREIGN ACCENTS

This chapter is used to express how accents are established, and why non-native English (foreign) speakers that have the same native language have similar accents when speaking English. The foreign accent similarities existing between these non-native speakers may potentially be used to identify the individual's native tongue and the country where they were raised.

The production of a foreign accent may be caused by many different factors, to include when, where, how, and why another language is learned. The theory used in this study is based on non-native English speakers learning English when they are well established speakers of their native language and are no longer children.

A. ACCENT PREMISE

The premise of limits on phonetic accuracy, [1, 2] may be simply stated as the old phrase, "You Can't Teach Old Dogs New Tricks". The speaker's native language is the source language, and the non-native spoken language is the target language. The phones of the speaker's native language are identified as L1 phones, and the phones of the target language are identified as L2 phones. Foreign accents may be caused from the production of sounds in the target language that are not used in the source language. Thus, the sounds in the target language that are not present in the source language will be the sounds most difficult to produce because these foreign sounds have never been used. This

production of accents premise is that if the target language has a sound that is not used in the speaker's source language, then the speaker will substitute an existing sound in the source language for the sound in the target language. If the source language sound is similar enough to be understandable, then the non-native speaker has no immediate incentive to improve on the pronunciation of the L2 phone. The production of similar accents from speakers with the same native language is then caused by the similar substitution of L1 phones for L2 phones. The identification of the L1 for L2 substitution is the key to recognizing a foreign accent. Note however that, not every non-native speaker learns a new language in the same way and not all L1 phones are pronounced the same. The seemingly simple task of identification of foreign accents is actually quite difficult to do automatically, and the difficulties in the identification process increase as the proficiency of the non-native speaker in the foreign language improves. The more phonemes that are not present in the speakers source language, the easier it should be to find accent possibilities. The idea is to start with the phonemes that are different from the source language however similar enough to cause a substitution of L1 for L2 phones, and then to look at the phonemes that had to be learned from scratch.

B. WORD LIST SELECTION

The word list selection is accomplished by identifying the L2 phones that are most difficult to pronounce for the foreign speakers. These phones may not be brand new phones, they may be target phones that are just close enough to existing phones in the speaker's source language so that a substitution of L1 for L2 phones seems harmless. A

brief interview of six native Brazilian speakers revealed the sounds that were most difficult to produce were the English phonemes /æ/, /fõ/, /ɝ/ and the sound created when pronouncing "rl" (as in "world" and "girl"). All these sounds are incorporated in the word list used. The word list is chosen using different phonemes in similar words so that when the similar words are spoken the only difference in the pronunciation is the phoneme of interest. Table 4 shows the word list used for the recordings, the vowel phonemes with the IPA, and the formant frequencies associated with each vowel phoneme.

TABLE 4 FOURTEEN-WORD LIST WITH VOWEL FORMANTS AND PHONEMES

Words	**F1**	**F2**	**International Phonetic**
world	490	1350	ɝ
men	530	1840	ɛ
sit	390	1990	I
tree	270	2290	i
man	660	1720	æ
being	N/A	N/A	iI
fifth	270	2290	I
zap	660	1720	æ
set	530	1840	ɛ
girl	490	1350	ɝ
seeing	N/A	N/A	iI
three	270	2290	i
sat	660	1720	æ
word	490	1350	ɝ

Only the voiced vowel phonemes are shown even though the entire word for each recording is AR modeled. Recall that differences in the resulting AR models of the full words, and the vowel phonemes contained in those words are very small (see Figure 7).

VI. PERFORMANCE MEASURES AND TESTING

This chapter presents performance tests which measure how well each speaker pronounces the selected set of words listed in Table 4 which contain particular phonemes, in comparison to a diversified reference group of native American English speakers. Five performance measures using AR models obtained from given words are used. AR models are produced using the entire word for each word on the word list in Table 4. The frequency region for the AR models is limited to the interval 0 to 2400 Hz, as described in Chapter III. As a result, the total number of points in each AR model sequence of order $P = 24$ is $N = 300$ which corresponds to 2400 Hz, given that 512 points are used to represent the AR frequency response. The five performance measures include; the Itakura distance [3, 8], the normalized cross-correlation coefficient and the modified normal cross-correlation coefficient [9], the log spectral distance [10], and a "bounds" measure defined in this study.

The list of fourteen words shown in Table 4 is repeated twice by each speaker which leads to a set of twenty-eight words per speaker. Each word is modeled using an AR model of order twenty-four. From the thirty-one native English speakers, sixteen are selected for an English speaking reference group. The remaining fifteen native English speaker's recordings are performance tested against the reference group. The Brazilian recordings are also performance tested against the native English speaking reference group. The test group consists of all the native English speakers not in the reference

group and all the non-native English speakers (Brazilians in this study). A reference AR model for each word is produced by calculating the mean of the sixteen selected AR models obtained from the native English speakers in the reference group. Figure 12 shows the reference AR model, highlighted with asterisks, from the AR models for a selected reference group of sixteen native English speakers for the word "girl". The reference model is used as the basis for all of the performance measures except the boundary measure. The following sections first describe each performance measure, and next explain how each AR modeled word in the twenty-eight word set is tested against a reference.

A. SYMMETRIZED ITAKURA DISTANCE

The Itakura distance enhances the effects of spectral differences due to the locations of the AR model peaks [3]. The AR model peaks indicate the formant frequencies present. The valleys of the AR model are not enhanced, therefore the errors from the differences in the valleys between the reference model and the tested speaker are not weighted as heavily as the differences in peaks. The formant frequencies are the frequencies of interest and here in determining the quality of the phoneme pronunciation or if a foreign accent is present.

The Itakura distance has been used extensively in speech applications [3]. It is not a metric which means it does not have the symmetry property. For example: if $v(\omega)$ is the spectral information corresponding to a speaker to be tested, and $Ref(\omega)$ is some reference model (obtained as the magnitude mean of the reference group), and $Itk(v(\omega),Ref(\omega))$ is

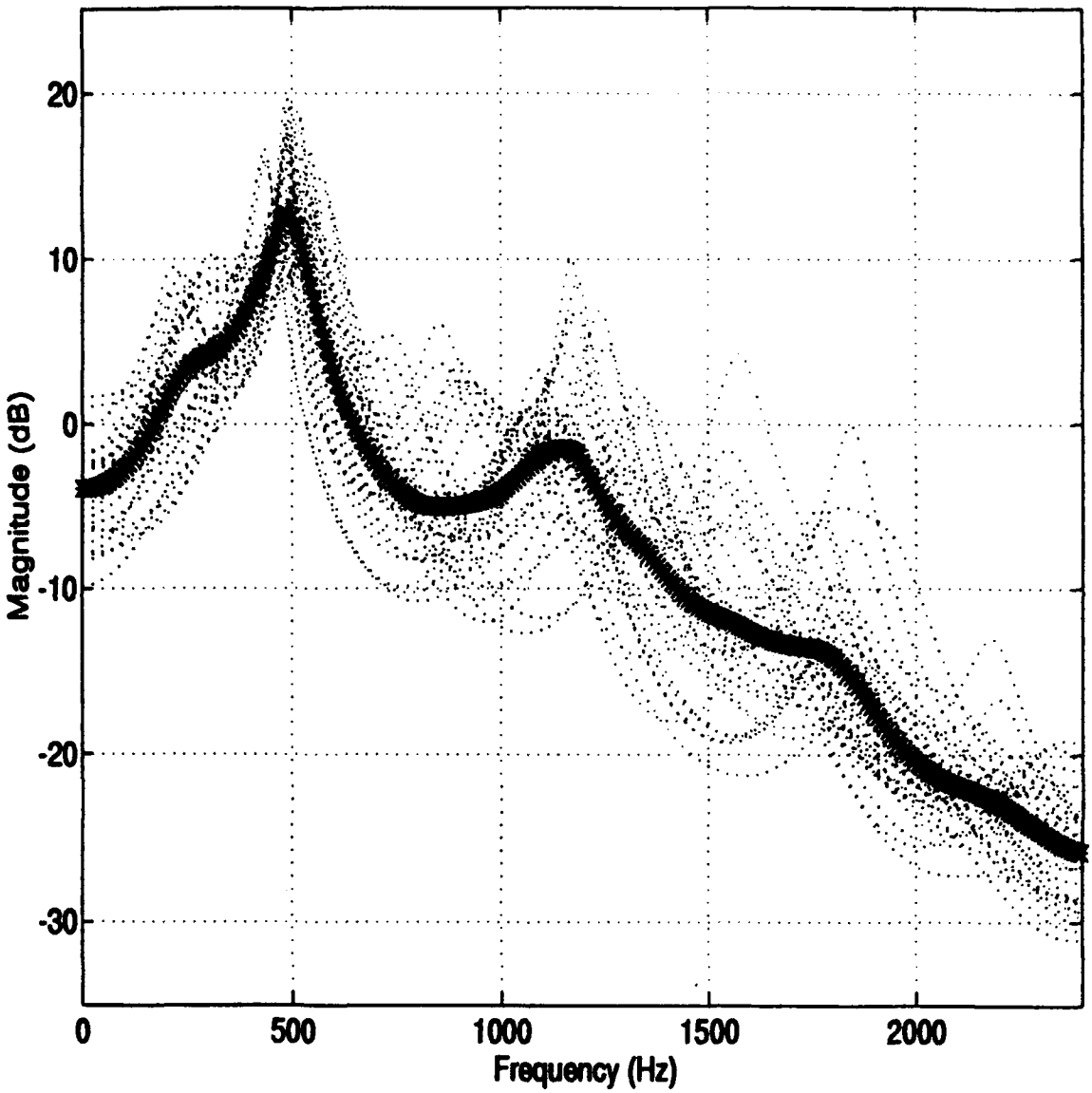


Figure 12: AR spectra obtained for the word "girl" for sixteen native male English speakers; resulting mean spectra (reference model) highlighted with asterisks.

defined as the Itakura distance between them then:

$$Itk(v(\omega),\ Ref(\omega) \neq Itk(Ref(\omega),\ v(\omega)).$$

To eliminate the above problem, the symmetrized Itakura distance measure is defined as [8]:

$$Itk(v, Ref) = ln\frac{1}{2\pi}\sqrt{\left(\int_{-\pi}^{\pi}\frac{v(\omega)}{Ref(\omega)}d\omega\right)\left(\int_{-\pi}^{\pi}\frac{Ref(\omega)}{v(\omega)}d\omega\right)}. \quad (7)$$

Equation (7) satisfies the symmetry property:

$$Itk(v(\omega),\ Ref(\omega) = Itk(Ref(\omega),\ v(\omega)).$$

The symmetrized Itakura distance does have the property that a measure from two identical AR models is zero, for example:

$$Itk(Ref, Ref) = 0 \text{ and } Itk(v,v) = 0, \quad (8)$$

therefore as the symmetrized Itakura distance between two spectra increases, the similarities between these two spectra decreases.

1. Application of the symmetrized Itakura distance

The AR model obtained from a speech signal $s(n)$ is expressed in terms of the AR coefficients defined in Chapter III as:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_P z^{-P} \text{ and the gain } b_o.$$

The spectrum $S(\omega)$ of $s(n)$ is obtained from the magnitude squared of the frequency response of the associated transfer function:

$$S(\omega) = |H(z)|^2_{z=e^{j\omega}} = \left|\frac{b_o}{A(z)}\right|^2_{z=e^{j\omega}}.$$

The MATLAB™ implementation of the symmetrized Itakura distance measure is presented in Appendix B.

2. Testing using the Itakura distance

A reference model tested against itself produces a measure equal to zero as shown

in equation (8). Every frequency response in the selected reference group is tested against a reference model (recall that the reference model is the mean of the selected AR spectra obtained from the reference group). The largest Itakura distance obtained from the reference model to all models contained in the reference group is labeled $Ref_{(max)}$. Next, each model contained in the test group is compared to the reference model and the resulting distance is compared to $Ref_{(max)}$. Every speaker in the test group that has an Itakura distance measure larger than $Ref_{(max)}$ is marked as a failure for the Itakura measure distance test corresponding to that word of the twenty-eight word set.

B. CROSS-CORRELATION COEFFICIENT

Two cross-correlation coefficients are used in this study; the normalized cross-correlation coefficient, which is referred to as cross-correlation-1, and the modified normalized cross-correlation coefficient measure, which is referred to as cross-correlation-2 [9]. Both cross-correlation measures use the reference model described in the introduction to this chapter and illustrated using the word "girl", as shown in Figure 12.

1. Normalized cross-correlation coefficient

The normalized cross-correlation coefficient cross-correlates the reference model with AR models to be tested and normalizes the results:

$$\rho_{rt} = \frac{\sum_{n=0}^{N-1} r(n)t(n)}{\sqrt{\sum_{n=0}^{N-1} r^2(n) \sum_{n=0}^{N-1} t^2(n)}},$$

where ρ_n is the normalized cross-correlation coefficient measure, $r(n)$ is the AR spectrum obtained for the reference model, $t(n)$ is the AR spectrum obtained for one word in the test group, and N is the number of frequency points considered for the test (for this study $N = 300$). Note that, ρ_n has a numerical value between zero and positive one. A numerical value of one $\rho_n = 1$ means that the two sequences $r(n)$ and $t(n)$ are identical, while a numerical value of zero $\rho_n = 0$ means that the two sequences $r(n)$ and $t(n)$ have zero percent correlation. The normalized cross-correlation coefficient measure determines the percent of correlation between the reference model and any test AR model.

2. Modified normalized cross-correlation coefficient

The modified normalized cross-correlation coefficient is defined the same as the normalized cross-correlation coefficient measure, except that before the procedure of cross-correlation, the mean of the reference model and each test AR model are removed. The range of possible numerical values for the modified normal cross-correlation coefficient measure is between negative and positive one. The case of identical sequences with a numerical value of one $\rho_n = 1$ still holds. For the case of no correlation, the numerical value would be zero $\rho_n = 0$.

3. Application of the cross-correlation coefficients

Cross-correlation-1 and cross-correlation-2 are implemented using the same procedure. For every word in the word list, each AR model is tested against a reference model using both cross-correlation-1 and cross-correlation-2. The selected reference

group is screened to determine the minimum numerical value for both cross-correlation tests. The minimum value from the reference group $\rho_{(min)}$ is compared against each numerical value $\rho_{(test)}$ calculated from the AR models in the test group. The MATLAB™ implementation of the cross-correlation coefficients is presented in Appendix C.

4. Testing using cross-correlation coefficients

For every speaker in the test group, the magnitudes of the cross-correlation measures are compared against the minimum value $|\rho_{(min)}|$ of the reference group. Every speaker in the test group may receive a failure for each time a recorded word has a cross-correlation numerical magnitude less than $|\rho_{(min)}|$. A total of four failures may be received for the cross-correlation tests for a single word since each word is recorded twice and both cross-correlation measures are used.

C. LOG SPECTRAL DISTANCE

The log spectral distance uses the reference model described in the introduction to this chapter and shown using the word "girl" for a selected reference group in Figure 12. The log spectral distance computes the sum of the difference between the frequency components of the AR spectrum, expressed in dB, obtained for the reference model and any of the components in the test group. The resulting log spectral distance expression is given by:

$$CB = \sum_{i=1}^{N} |\log(AR_{M,}) - \log(AR_{T,})|,$$

where CB is the log spectral distance, AR_{Mi} is the spectrum value at frequency location $\frac{2\pi j}{512}$ for the reference model, AR_{Ti} is the spectrum value at frequency location $\frac{2\pi j}{512}$ for a component of the test group, and the parameter N is the number of frequency points considered in this study.

1. Application of the log spectral distance

The log spectral distance is used to test each AR model in the test group against the reference model for each word recorded. The MATLAB™ implementation of the log spectral distance is presented in Appendix D.

2. Testing with the log spectral distance

Every AR modeled word is tested against the reference model, including every word in the reference group. The maximum log spectral distance calculated for the reference group is used to compare each calculated log spectral distance from the test group. For each log spectral distance from the test group that is greater than the maximum log spectral distance obtained for the reference group, a failure is marked for that speaker. Each speaker may fail the log spectral distance twice for each word on the selected word list (Table 4), since each word is recorded twice.

D. "BOUNDS" MEASURE

The "bounds" measure is used to identify differences in AR model frequency locations or AR model shapes indicating different sounds. The reference model described in the introduction of this chapter is not used for the "bounds" measure. The AR model

spectrum magnitude upper and lower values are the reference bounds for this measure. The reference upper bound is obtained by taking the maximum magnitude for each frequency component of the AR spectra from the native English speaking reference group. Similarly, the reference lower bound is obtained by taking the minimum magnitude for each frequency component of the AR spectra from the native English speaking reference group. Figure 13 shows a selected native English speaking reference group of AR models for the word "girl" with the bounds highlighted with asterisks.

1. Application of the "bounds" measure

The reference bounds are computed for each word on the word list (Table 4), and then each AR model of the native English and non-native English speaking test groups are tested against the bounds. The MATLAB™ implementation of the "bounds" measure is presented in Appendix E.

2. Testing using the "bounds" measure

The AR modeled words are tested by determining the percentage of each AR modeled word from the test group that is outside the bounds. The speaker is marked with a failure for the "bounds" measure for each word when for both times the particular word is recorded, five percent of the magnitude of the frequency response of the AR model is located outside the reference bounds. Experimentally, five percent of the magnitude of the frequency response of the AR model outside the reference boundary proved to be satisfactory for the list of words considered. Each speaker may only receive one failure of the "bounds" measure for each word on the selected word list (Table 4).

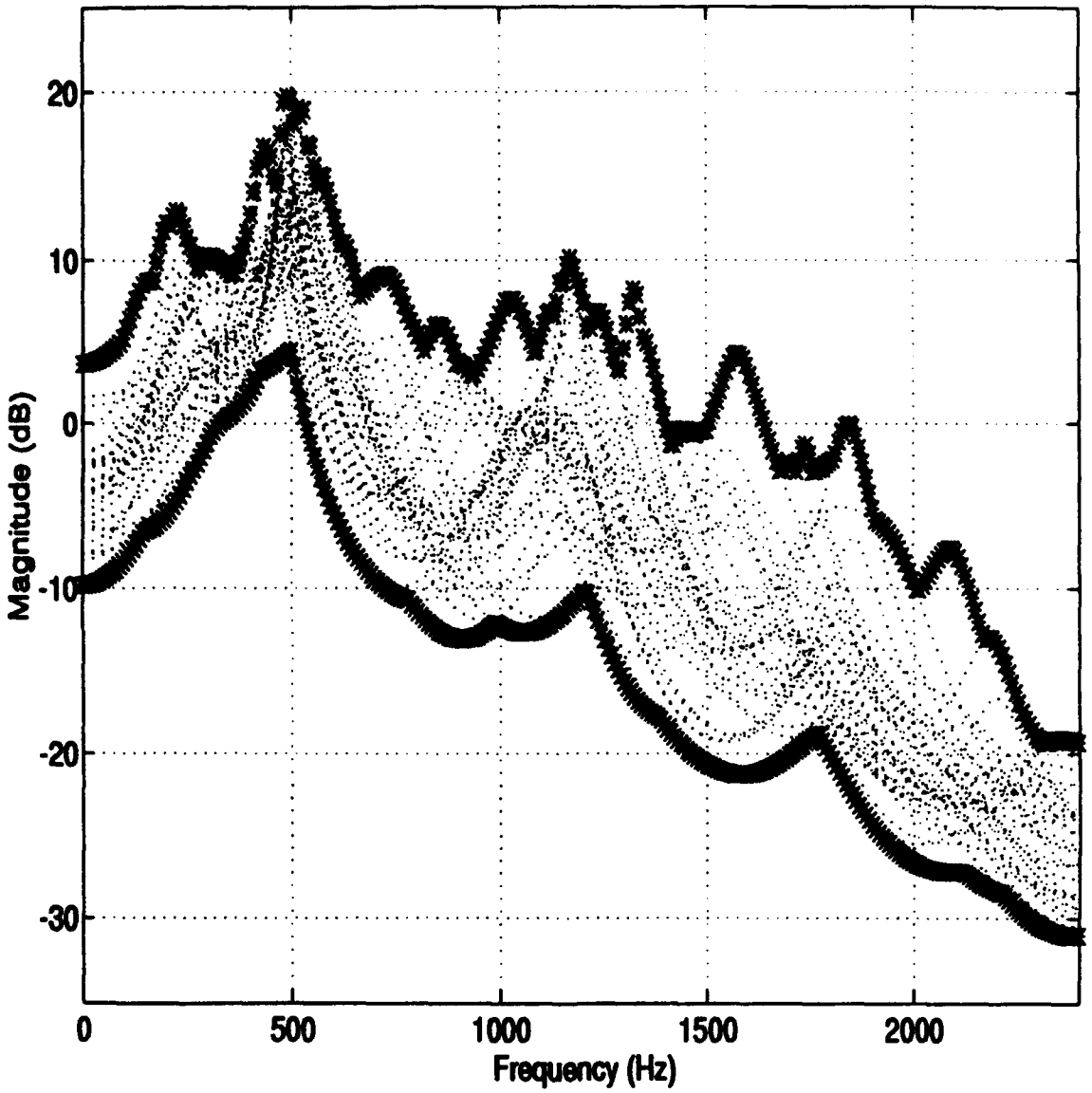


Figure 13: AR spectra obtained for the word "girl" for sixteen native male English speakers; resulting reference "bounds" highlighted with asterisks.

VII. MODELS AND TEST RESULTS

The goal of performance testing is to ensure results are achieved from a robust set of tests with a detailed method of analysis. The five performance measures used here to determine whether a given speaker is or is not a native American English speaker have been described earlier in Chapter VI. The results of the five performance measures are combined to determine if the tested speaker is a native English speaker, a non-native (foreign) English speaker or marginal. A marginal rating means that the speaker could be a native or a non-native English speaker. A non-native English speaker with a minor accent could potentially pass as a native English speaker and conversely, a native English speaker with an anomaly in his phoneme production or with a regional native English accent not sufficiently contained in the reference group could fail as a native English speaker.

A. REFERENCE MODELS

Reference models are used to determine how well speakers in the test group pronounce phonemes contained in full words. Recall that the reference model is obtained from a selected reference group of sixteen speakers from the thirty-one native English speakers recorded. To eliminate potential bias in the results, six reference groups are selected and used to obtain a reference model and reference bounds for the "bounds" measure. Each reference model and reference bound is then used to test the associated test group. Recall that the test group consists of the native English speakers not included in the reference group and the non-native (Brazilian) English speakers.

The native English speakers are numbered in the order they are recorded. The numbering scheme for the native English speakers is S1, S2, · · ·, S31. The Brazilian speakers are labeled similarly where B1 is the first Brazilian non-native English speaker recorded, and B6 is the last Brazilian non-native English speaker recorded (B1,B2, · · ·, B6). The first reference group, labeled RG1 for simplicity, used in this study consists of the first sixteen native English speakers recorded, i.e., RG1 = (S1, S2, · · ·, S16). The next fifteen native English speakers recorded as well as the Brazilians recorded make up the test group, labeled TG1 = (S17, S18, · · ·, S31, B1, B2, · · ·, B6). The second reference group consists of the last sixteen native English speakers with respect to the order in which they were recorded. RG2 = (S16, S18, · · ·, S31), and the associated test group is TG2 = (S1, S2, · · ·, S15, B1, B2, · · ·, B6). The third reference group consists of all the odd numbered native English speakers with respect to the order in which they were recorded RG3 = (S1, S3, · · ·, S31), and the associated test group is TG3 = (S2, S4, · · ·, S30, B1, B2, · · ·, B6). Additional reference and test groups are selected randomly and they include:

- RG4 = (S1, S2, S3, S8, S12, S13, S16, S18, S20, S21, S22, S23, S24, S26, S28, S29), with TG4 = (S4, S5, S6, S7, S9, S10, S11, S14, S15, S17, S19, S25, S27, S30, S31, B1, B2, · · ·, B6);
- RG5 = (S4, S8, S9, S10, S13, S14, S16, S17, S18, S21, S25, S26, S27, S28, S29, S30), with TG5 = (S1, S2, S3, S5, S6, S7, S11, S12, S15, S19, S20, S22, S23, S24, S31, B1, B2, · · ·, B6);

- RG6 = (S1, S2, S3, S4, S5, S6, S7, S12, S14, S19, S21, S23, S25, S26, S29, S30), with TG6 = (S8, S9, S10, S11, S13, S15, S16, S17, S18, S20, S22, S24, S27, S28, S31, B1, B2, · · ·, B6).

The native English speakers are from many areas of the United States, the states included are: California, Connecticut, Florida, Kansas, Maryland, Massachusetts, Minnesota, Mississippi, Missouri, New Jersey, New York, North Dakota, Ohio, Oregon, South Carolina, Virginia, and Wisconsin.

B. THRESHOLDS

Chapter VI defines individual word failures for each performance measure. Recall that all of the performance measures except the "bounds" measure record failures for each word individually and since each word on the word list in Table 4 is recorded twice a set of twenty-eight words are potential failures. The "bounds" measure uses a method which combines the results obtained from both recordings of a given word, leaving a potential of fourteen word failures. Next, thresholds are required to determine how many of these word failures constitute a failure for each performance measure. The performance measure thresholds are related to the number of words on the word list in Table 4. Thresholds are also required to determine the number of performance measure failures that establish a rating of each speaker as; native English speaker, non-native English speaker, or marginal. The rating thresholds are not dependent on the number of words on the word list in Table 4. Note that all thresholds are obtained heuristically through experimentation. Ideally, the results should show that all native English speakers tested

against any of the native English speaker reference groups are determined to be native English speakers, and that all non-native English speakers tested against any native English speaking reference group are determined to be non-native English speakers. Table 5 shows the thresholds set for the number of word failures that are required for each performance measure to be considered a failure.

TABLE 5 THRESHOLDS FOR PERFORMANCE MEASURE FAILURE USING THE FOURTEEN-WORD LIST

TEST	ITK	UR	RR	CB	BND
THRESHOLD	3	3	3	3	4

The test names are abbreviated such that; ITK is the symmetrized Itakura distance , UR is the normalized cross-correlation coefficient, RR is the modified normalized cross-correlation coefficient, CB is the log spectral distance, and BND is the "bounds" measure.

Rating thresholds are determined experimentally and are not related to the number of words on the word list in Table 4. The thresholds for the number of performance measure failures that establish a rating of either native English speaker, marginal, or non-native English speaker, are listed in Table 6.

TABLE 6 THRESHOLDS FOR RATINGS

RATING	THRESHOLD
Native	≤ 2
Marginal	3
Non-Native	≥ 4

The number of performance measure failures is arrived at by counting the number of failures that meet the thresholds established in Table 5, and adding to that number 0, if 0 is recorded for any one performance measure, and adding 2 if there are no zeros recorded for any one performance measure. This offset was chosen to enhance the results for a perfect score for any one performance measure. The results of this study (using all reference groups, with the reduced word list explained later in this chapter) show that 89% of all native English speakers score 0 for one of the five performance measures when tested. Table 7 shows an example of how the ratings are calculated. The columns in Table 7 labeled; ITK, UR, RR, CB, and BND contain the number of failures of each performance measure using the rules for failure established in Chapter VI. The number of performance measure failures established by the thresholds listed in Table 5 are counted and recorded in the column labeled PMF (Performance Measure Failures). The column labeled ZS (Zero Scored) reflects the results of a zero recorded for any performance measure (zero for a zero recorded and two for no zero recorded). The column labeled Total is the total of the two columns labeled; PMF and ZS. The column labeled Rating is scored by reviewing the numbers listed in the Total column and using the thresholds listed in Table 6. A speaker is given a rating of: N for native English speaker, F for non-native (Foreign) English speaker, and M for marginal.

TABLE 7 EXAMPLE RATING CALCULATIONS

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S1	1	0	0	4	3	1	0	1	N
S2	1	1	2	1	2	0	2	2	N
S3	0	0	0	0	2	0	0	0	N
B1	3	5	7	5	6	5	2	7	F
B2	0	4	3	1	5	3	0	3	M

C. TEST RESULTS

The test results shown in the following tables have the same format as the example shown in Table 7. The results for test group 1 (TG1) with reference group 1 (RG1) are presented in Table 8. The results from TG1 show that 100% of all the non-native English speakers tested received a rating of foreign, none of the non-native English speakers received a rating of marginal or native, 73% of all the native English speakers tested received a rating of native, 20% of the native English speakers received a rating of foreign, and one native English speaker corresponding to 7% received a rating of marginal. All of the other five test group results are calculated in the same way as for the example in Table 7 and for the results shown in Table 8.

TABLE 8 RESULTS FOR TG1 WITH RG1 (FOURTEEN-WORD LIST)

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S17	0	0	0	0	2	0	0	0	N
S18	2	2	3	0	3	0	0	0	N
S19	0	0	0	0	0	0	0	0	N
S20	0	0	0	0	0	0	0	0	N
S21	1	2	2	1	1	0	2	2	N
S22	0	3	0	1	1	1	0	1	N
S23	2	1	3	0	3	1	0	1	N
S24	1	1	1	2	2	0	2	2	N
S25	3	0	0	1	2	1	0	1	N
S26	1	0	0	1	2	0	0	0	N
S27	0	0	0	0	1	0	0	0	N
S28	1	7	3	3	6	4	2	6	F
S29	6	3	4	4	4	5	2	7	F
S30	11	13	12	8	9	5	2	7	F
S31	5	0	2	4	7	3	0	3	M
B1	3	1	1	5	2	2	2	4	F
B2	4	4	5	4	5	5	2	7	F
B3	2	2	2	4	5	2	2	4	F
B4	7	9	7	7	6	5	2	7	F
B5	3	3	3	5	2	4	2	6	F
B6	2	7	4	2	5	3	2	5	F

Table 9 summarizes the test results obtained for all combinations considered. The abbreviated headings for each column in Table 9 are: PNRN (Percentage of Native speakers Rated as Native speakers), PFRF (Percentage of Foreign speakers Rated as Foreign speakers), PNRM (Percentage of Native speakers Rated as Marginal), PFRM (Percentage of Foreign speakers Rated as Marginal), PNRF (Percentage of Native

speakers Rated as Foreign speakers), PFRN (Percentage of Foreign speakers Rated as Native speakers), and STD (STandard Deviation).

TABLE 9 SUMMARY OF TEST RESULTS FOR FOURTEEN-WORD LIST

Test	PNRN	PFRF	PNRM	PFRM	PNRF	PFRN
TG1	73 %	100 %	7 %	0 %	20 %	0 %
TG2	73 %	67 %	0 %	17 %	27 %	16 %
TG3	87 %	67 %	0 %	33 %	13 %	0 %
TG4	47 %	83 %	7 %	0 %	46 %	17 %
TG5	93 %	58 %	0 %	17 %	7 %	25 %
TG6	93 %	67 %	0 %	0 %	7 %	33 %
Mean	78 %	74 %	2 %	11 %	20 %	15 %
1 STD	16 %	14 %	3 %	12 %	14 %	12 %

Results shown in Table 9 for all tests using six different reference groups indicate high levels of missclassification. Thus, the word list shown in Table 4 must be restricted to the words which are considered by the non-native English speaker as the most difficult ones to pronounce.

The fourteen-word list from Table 4 is reduced to five words: "man", "zap", "girl", "seeing", and "word". Recall that the performance measure failure thresholds are dependent on the number of words contained on the word list, therefore the thresholds for the reduced word list are also reduced. Table 10 shows the performance measure failure thresholds for the reduced five-word list.

TABLE 10 THRESHOLDS FOR PERFORMANCE MEASURE FAILURE (5 WORD)

TEST	ITK	UR	RR	CB	BND
THRESHOLD	2	2	2	2	3

The thresholds that determine the rating of a speaker remain constant and are listed in Table 6.

Table 11 through Table 16 show the results for TG1 through TG6 using the reduced five-word list. Table 11 contains the results for TG1 with RG1.

TABLE 11 RESULTS FOR TG1 WITH RG1 USING THE FIVE-WORD LIST

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S17	0	0	0	0	2	0	0	0	N
S18	0	2	1	0	3	2	0	2	N
S19	0	0	0	0	0	0	0	0	N
S20	0	0	0	0	0	0	0	0	N
S21	0	1	1	1	0	0	0	0	N
S22	0	1	0	1	1	0	0	0	N
S23	0	0	1	0	0	0	0	0	N
S24	0	0	0	2	1	1	0	1	N
S25	1	0	0	1	0	0	0	0	N
S26	0	0	0	1	1	0	0	0	N
S27	0	0	0	0	1	0	0	0	N
S28	0	3	1	1	3	2	0	2	N
S29	3	3	3	3	1	4	2	6	F
S30	3	4	4	1	3	4	2	6	F
S31	1	0	0	2	3	2	0	2	N
B1	1	1	1	3	1	1	2	3	M
B2	3	4	5	4	3	5	2	7	F
B3	2	2	2	4	4	5	2	7	F
B4	6	6	6	5	3	5	2	7	F
B5	2	2	2	4	2	4	2	6	F
B6	2	3	3	2	2	4	2	6	F

Table 12 contains the results for TG2 with RG2, Table 13 contains the results for TG3 with RG3, Table 14 contains the results for TG4 with RG4, Table 15 contains the results for TG5 with RG5, and Table 16 contains the results for TG6 with RG6.

TABLE 12 RESULTS FOR TG2 WITH RG2 USING THE FIVE-WORD LIST

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S1	1	0	0	1	0	0	0	0	N
S2	1	0	0	0	2	0	0	0	N
S3	0	0	0	0	1	0	0	0	N
S4	3	1	4	0	1	2	0	2	N
S5	4	0	1	0	2	1	0	1	N
S6	0	0	0	0	1	0	0	0	N
S7	0	1	0	0	2	0	0	0	N
S8	0	0	0	0	1	0	0	0	N
S9	1	1	0	0	0	0	0	0	N
S10	0	1	0	0	1	0	0	0	N
S11	4	1	0	0	2	1	0	1	N
S12	1	0	0	0	2	0	0	0	N
S13	1	0	0	0	0	0	0	0	N
S14	4	3	4	1	2	3	2	5	F
S15	0	0	0	0	0	0	0	0	N
B1	1	0	1	1	2	0	0	0	N
B2	7	3	4	4	3	5	2	7	F
B3	2	2	2	2	1	4	2	6	F
B4	5	5	5	6	3	5	2	7	F
B5	5	2	2	3	3	5	2	7	F
B6	4	2	1	0	1	2	0	2	N

TABLE 13 RESULTS FOR TG3 WITH RG3 USING THE FIVE-WORD LIST

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S2	0	0	0	2	1	1	0	1	N
S4	1	2	2	0	0	2	0	2	N
S6	0	0	0	0	0	0	0	0	N
S8	0	0	0	0	1	0	0	0	N
S10	0	1	0	1	2	0	0	0	N
S12	0	0	0	2	2	1	0	1	N
S14	1	2	2	1	3	3	2	5	F
S16	0	0	0	1	2	0	0	0	N
S18	0	3	0	0	1	1	0	1	N
S20	0	0	0	0	0	0	0	0	N
S22	0	1	0	1	1	0	0	0	N
S24	0	0	0	0	0	0	0	0	N
S26	0	0	0	1	1	0	0	0	N
S28	0	1	0	0	1	0	0	0	N
S30	2	4	5	4	3	5	2	7	F
B1	1	1	1	1	2	0	2	2	N
B2	3	4	4	5	3	5	2	7	F
B3	2	2	2	3	1	4	2	6	F
B4	6	5	5	5	4	5	2	7	F
B5	2	2	2	2	2	4	2	6	F
B6	0	2	2	1	1	2	0	2	N

TABLE 14 RESULTS FOR TG4 WITH RG4 USING THE FIVE-WORD LIST

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S4	3	1	3	0	0	2	0	2	N
S5	3	0	2	0	1	2	0	2	N
S6	0	0	0	0	1	0	0	0	N
S7	1	1	1	0	1	0	0	0	N
S9	1	1	0	0	0	0	0	0	N
S10	1	1	2	1	1	1	2	3	M
S11	3	2	2	3	2	4	2	6	F
S14	5	2	3	2	2	4	2	6	F
S15	1	0	0	0	0	0	0	0	N
S17	1	0	0	0	0	0	0	0	N
S19	0	0	0	0	0	0	0	0	N
S25	0	0	0	0	0	0	0	0	N
S27	0	0	0	0	0	0	0	0	N
S30	6	6	5	7	3	5	2	7	F
S31	4	0	0	5	2	2	0	0	N
B1	3	1	1	3	3	3	2	5	F
B2	6	3	5	5	3	5	2	7	F
B3	2	2	2	2	1	4	2	6	F
B4	7	5	6	7	3	5	2	7	F
B5	5	2	2	4	3	5	2	7	F
B6	5	3	3	3	1	4	2	6	F

TABLE 15 RESULTS FOR TG5 WITH RG5 USING THE FIVE-WORD LIST

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S1	1	0	0	2	1	1	0	1	N
S2	0	0	0	0	1	0	0	0	N
S3	0	0	0	0	0	0	0	0	N
S5	0	0	0	0	1	0	0	0	N
S6	0	0	0	0	0	0	0	0	N
S7	0	0	0	0	1	0	0	0	N
S11	3	1	0	2	3	3	0	3	M
S12	0	0	0	0	2	0	0	0	N
S15	0	0	0	0	0	0	0	0	N
S19	0	0	0	0	0	0	0	0	N
S20	0	0	0	0	0	0	0	0	N
S22	0	0	0	0	0	0	0	0	N
S23	0	0	0	0	0	0	0	0	N
S24	0	0	0	1	0	0	0	0	N
S31	1	0	0	3	3	2	0	2	N
B1	2	1	1	3	2	2	2	4	F
B2	4	3	3	4	4	5	2	7	F
B3	2	2	2	2	1	4	2	6	F
B4	5	4	4	4	2	4	2	6	F
B5	3	2	2	4	5	5	2	7	F
B6	1	1	1	0	2	0	0	0	N

TABLE 16 RESULTS FOR TG6 WITH RG6 USING THE FIVE-WORD LIST

Speaker	ITK	UR	RR	CB	BND	PMF	ZS	Total	Rating
S8	0	0	0	0	0	0	0	0	N
S9	0	0	0	0	0	0	0	0	N
S10	0	0	0	2	2	1	0	1	N
S11	3	1	1	2	2	2	2	4	F
S13	0	0	0	0	1	0	0	0	N
S15	0	0	0	0	0	0	0	0	N
S16	0	0	0	1	0	0	0	0	N
S17	0	0	0	0	1	0	0	0	N
S18	0	0	0	0	0	0	0	0	N
S20	0	0	0	0	0	0	0	0	N
S22	0	1	0	0	0	0	0	0	N
S24	0	0	0	1	1	0	0	0	N
S27	0	0	0	0	0	0	0	0	N
S28	0	0	0	0	1	0	0	0	N
S31	0	0	0	2	2	1	0	1	N
B1	1	0	1	5	3	2	0	2	N
B2	2	3	3	3	3	5	2	7	F
B3	2	2	2	2	1	4	2	6	F
B4	5	5	5	6	4	5	2	7	F
B5	2	2	2	3	2	4	2	6	F
B6	2	1	1	2	3	3	2	5	F

Table 17 summarizes all six test group results using the reduced five-word list, and this recapitulation shows more convincing results. The abbreviations for the column headings in Table 17 are the same as for Table 9 (p. 49).

TABLE 17 SUMMARY OF TEST RESULTS FOR THE FIVE-WORD LIST

Test Group	PNRN	PFRF	PNRM	PFRM	PNRF	PFRN
TG1	87 %	83 %	0 %	17 %	13 %	0 %
TG2	93 %	67 %	0 %	0 %	7 %	33 %
TG3	87 %	67 %	0 %	0 %	13 %	33 %
TG4	73 %	100 %	7 %	0 %	20 %	0 %
TG5	93 %	83 %	7 %	0 %	0 %	17 %
TG6	93 %	83 %	0 %	0 %	7 %	17 %
Mean 1 STD	88 % 7 %	80.5 % 11 %	2 % 3 %	2.8 % 6 %	10 % 6 %	16.7 % 13 %

Comparing results from this reduced word list with those listed in Table 9 (p. 49) obtained when using the fourteen-word list shows that a distinct improvement in classification performance has been obtained, however PFRN shows a small degradation. This comparison also shows that the selection of the word list is one of the key factors in the automatic classification of native versus non-native speakers.

The MATLAB™ implementation of the results is presented in Appendix F.

VIII. CONCLUSIONS

The goal of accent recognition investigated in this thesis is to automatically detect non-native (foreign) English speakers as foreign, and native American English speakers as native using AR modeling. The processing techniques are simple to implement and data preparation is automated. The entire process from spoken word to rating of speaker can be automated for practical use.

This thesis considers the use of a few single syllable words common in daily speech, and focuses on one group of non-native English speakers, with the notion that the techniques used for accent detection may be extended to recognize non-native English speakers from many languages. The non-native English speakers selected for this study are all Brazilian students attending the Naval Postgraduate School. The word list used is made up of words that are difficult for native Brazilians to pronounce. The native English speakers used in this study are originally from various regions of the United States and are all military servicemen which limits regional accent due to the many areas of their travels and residences.

Results show that an average of 88 % of all native speakers tested are rated as native, and that an average of 80.5 % of all foreign speakers tested are rated as foreign. Six different reference groups of sixteen native English speakers are separately used to test fifteen native and six non-native English speakers. The robustness of the techniques is

improved by using various reference groups and maintaining the ability to produce similar results.

The results produced by this study are encouraging as they show that it may be possible to detect foreign accents. However, these results may be improved by; choosing better words for the performance tests, maintaining a cleaner environment for recording, and adding a time varying analysis technique to the performance measures. First, choosing better words for the performance test could improve results as word selection is critical to achieving accent recognition. Words may exist that have more consistency among native English speakers and cause more variances from the reference groups for foreign speakers, which would produce better results overall. Second, maintaining a cleaner environment for recording may provide higher accuracy for the AR models and emphasize the differences between native and non-native English speakers, which would improve the classification process. Finally, adding a time varying analysis technique to the performance measures may enhance the results by better showing the differences in the pronunciation of long vowels (diphthongs). The difficulties encountered in these procedures come in the form of relating different speakers pronouncing the same word over different duration's of time, and additional processing such as Dynamic Time Warping is then needed to align the spoken words. Another alternative would be to compare spectrograms (three dimensional spectra), and to compare the time sequencing of the frequencies present. One of the phenomenon discovered during this research is that the Brazilian speakers involved in the study pronounce diphthongs in a time increment that

does not support the sounds required. For example, they pronounce long vowel sounds too fast. However, it is difficult to match a native English speaker who speaks quickly with a Brazilian who mispronounces sounds by pronouncing them over too short a period of time.

An additional approach involving Cepstral analysis was investigated. However, we noted that it did not produce satisfying results for the tests designed for the AR models.

APPENDIX A MATLAB™ IMPLEMENTATION OF AR SPECTRA

```
% calcul.m, Calculates the AR spectra and stores them in matrix form
% calcul.m calls function arcorfmp.m (MATLAB™ implementation follows cal.m)

% Inputs
% g is the digitized word sequence to be AR modeled
% N is the length of the AR sequence output
% fs is the sampling frequency of the digitization
% P is the order of the AR model
% E is the number of speakers to be modeled
% [g]e# is the digitized word sequence for native English speaker #
% [g]e#a is the digitized word sequence for native English speaker # second recording of
%the same word
% [g]# is the digitized word sequence for non-native English speaker #
% [g]#a is the digitized word sequence for non-native English speaker # second
%recording of the same word
g = input('Enter the name of the word to be modeled:','s');
N = 512;
fs = 8192;
P = 24;
E = 31;

% Calculate the AR spectra for each native English speaker
for m = 1:E
    nn = num2str(m);
    [MHz] = arcorfmp(eval([g,'e',nn]));
    MHze(:,m) = MHz;
end
% Calculate the AR spectra for the second recording of the same word
for ma = 1:E
    nna = num2str(ma);
    [MHz] = arcorfmp(eval([g,'e',nna,'a']));
    MHzea(:,ma) = MHz;
end
% Create matrix of AR models for English speakers
AA = [MHze,MHzea];

% Clear variables
for m = 1:E
mm = num2str(m);
    clear (['beinge',mm]);clear (['fifthe',mm]);clear (['girle',mm]) clear (['mane',mm]);
    clear (['mene',mm]);clear (['sate',mm]) clear (['seeinge',mm]);clear (['sete',mm]);
```

```
    clear (['site',mm]);clear (['threee',mm]);clear (['treee',mm]);clear (['worde',mm]);
    clear (['worlde',mm]);clear (['zape',mm]); clear (['beinge',mm,'a']);
    clear(['fifthe',mm,'a']);clear (['girle',mm,'a']);clear (['zape',mm,'a'])
    clear (['mane',mm,'a']);clear (['mene',mm,'a']);clear (['sate',mm,'a'])
    clear (['seeinge',mm,'a']);clear (['sete',mm,'a']);clear (['site',mm,'a'])
    clear (['threee',mm,'a']);clear (['treee',mm,'a']);clear (['worde',mm,'a'])
    clear (['worlde',mm,'a']);
end; clear m mm E fs P N clne

%  Save each words AR spectra in a matrix AA
save (['AA',g])

%-------------------------------------------------------------------------------------------%
%  Called function
%-------------------------------------------------------------------------------------------%
%  arcf.m (function), Calculates the AR spectra from the digitized recordings
%  AR model using the autocorrelation method and ar_corr.m from the Naval
%  Postgraduate School SPC toolbox [11]
%  Inputs
%  data is the digitized recordings of each word separately
%  P is the order of the AR model desired
%  N is the length of the frequency response desired

function [MHz,xax,bo,a,data] = arcorfmp(data);
P = 24; N = 512;

%  Normalize the data
datamm = data - mean(data);
load B100;
fdata1 = filter(B100,1,datamm);
load lpf;
fdata = filter(Bb,Aa,fdata1);
fdatamm = fdata - mean(fdata);
fdatammn = fdatamm ./(sqrt(fdatamm'*fdatamm));
data = fdatammn;

%  Calculate the AR model coefficients and gain
[a,bo,s,R] = ar_corr(data,P);

%  Calculate the frequency response of the AR coefficients with gain bo
Hz = freqz(bo,a,N);
%  Calculate the power of the frequency response in dB
MHz = 20*log10(abs(Hz));
```

APPENDIX B MATLAB™ IMPLEMENTATION OF THE ITAKURA DISTANCE

```
% Itk.m, Itakura distance
% This program calculates the Itakura distance for a matrix of input AR spectra TG with
%respect to the reference model reff

% Inputs
% AA is the matrix of AR spectra from the reference group
% TG is the natrix of AR spectra from the test group and the reference model
% N is the length of the AR spectra considered
% nS is the number of speakers in the test group

% Calculate the reference model
reff = mean(AA');

% Calculate the frequency response from the AR spectra for the reference model
Sr = 1 ./(10 .^(reff ./ 10));
Sr = Sr(:);

% Calculate the Itakura distance
%%    Test group and reference model
for u = 1:2*nS+1;
S = 1 ./ (10 .^(TG(:,u) ./10));
dSdSr(u) = log(sum(Sr ./ S)) + log(sum(S ./ Sr)) + log(1/N/N);
end
%%    Reference group
for ue = 1:32
Srr = 1 ./(10 .^(AA(:,ue) ./10));
dSrdSr(ue) = log(sum(Sr ./ Srr)) + log(sum(Srr ./ Sr)) + log(1/N/N);
end

% Check, measure the reference model against itself dchk should equal zero
Schk =(1 ./(10 .^(reff ./10)))';
dchk = log(sum(Sr ./ Schk)) + log(sum(Schk ./ Sr)) + log(1/N/N);
```

APPENDIX C MATLAB™ CROSS-CORRELATION COEFFICIENTS

```
% Code implementation obtained from [9]

% ccdist.m     last modified 3/10/94                    MPF
% computes variuos distances between the AR spectra
% rr is the normalized cross correlation (no DC component present)
% ur is the normalized cross correlation (includes potential DC effects)

% Inputs
% AA is the matrix of AR spectra for the reference group
% TG is the matrix of AR spectra for the test group
%  n is the number of speakers in the test group

ref0=AA';
x=TG';

% Compute reference model
ref=mean(ref0);
[n,b]=size(x);

% Compute the modified reference model for the modified cross-correlation coefficient
refn=ref-mean(ref);

% Sum over each col. to get norm. ref.
refn=refn+eps*ones(size(refn));

% Compute the cross-correlation coefficients
for i=1:n
    xeps(i,:)=x(i,:)+eps*ones(size(x(i,:)));
    ur(i)=ref*x(i,:)'/(sqrt(ref*ref'*xeps(i,:)*xeps(i,:)'));
    y(i,:)=x(i,:)-mean(x(i,:));
    yeps(i,:)=y(i,:)+eps*ones(size(y(i,:)));
    rr(i)=refn*y(i,:)'/(sqrt(refn*refn'*yeps(i,:)*yeps(i,:)'));
    x(i,:)=xeps(i,:)/sum(xeps(i,:));
end
```

APPENDIX D MATLAB™ IMPLEMENTATION OF THE LOG SPECTRAL

```
% CB.m, Log Spectral Distance

% Inputs
% AA is the matrix of AR spectra for the reference group
% TG is the matrix of AR spectra for the test group
% nS is the number of speakers in the test group

% Compute the reference model
Ref = mean(AA')';

% Compute the log spectral distance
%%    Test group
for n = 1:(2*nS) + 1
    diff(:,n) = TG(:,n) - Ref;
    d(n) = sum(abs(diff(:,n)));
end
%%    Reference group
for m = 1:32
    diffe(:,m) = AA(:,m) - Ref;
    de(m) = sum(abs(diffe(:,m)));
end
```

APPENDIX E MATLAB™ IMPLEMENTATION OF THE BOUNDS MEASURE

```
% Bounds.m, Bounds measure

% INPUTS
% N = length of the AR spectra sequences considered
% AA is a matrix of AR spectra from the reference group
% TG is a matrix of AR spectra from the test group
% Tg is the number of speakers in the test group
Tg = nS;

% Calculate the Bounds
for n = 1: N
    lb(n,:) = min(AA(n,:));
    ub(n,:) = max(AA(n,:));
end

% Calculate the Percentage of AR spectra that is outside of the bounds
for ep = 1:2*Tg+1
    Ep = TG(:,ep);
    ebu(:,ep) = ub - Ep;
    ebl(:,ep) = Ep - lb;
    ce(ep,:) = size(find(ebu(:,ep)<0 | ebl(:,ep)<0));
end
cce = ce(:,1);

for ne = 1:Tg
    chke = ce(ne);
    if chke ~= 0
        faile(ne) = ne;
end,end

espk = ce(:,1);
for fne = 1:length(espk);
    pespk(fne) = 100*(ce(fne)/(2*N));
end

% Calculate the speakers that are outside the bounds by more than five percent
Tsg = zeros(size(pespk));
Eng_test_bnd = find(pespk>5);
Tsg(Eng_test_bnd) = ones(size(Eng_test_bnd));
```

```
% Calculate the Speakers that are outside the bounds for more than five percent for both
%times a word has been recorded.
for n = 1:Tg
    if  Tsg(n) == 1
        if  Tsg(n) == Tsg(n+15) ;
            TSG(n) = n;
        else
            TSG(n) = 0;
        end
    else
        TSG(n) = 0;
    end
end

clear n

% Fail_Bnd is the results of a failed bounds measure
for n = 1:Tg
    Fail_Bnd = find(TSG(1:Tg)>0);
end
```

APPENDIX F MATLAB™ IMPLEMENTATION OF THE RESULTS

```
% Results.m, Results calculation

% Inputs
% N is the length of the AR spectra considered
% AA is the matrix of the AR spectra for all native English speakers
% BB is the matrix of the AR spectra for all non-native English speakers
% TG is the matrix of the AR spectra for the test group
% ff is the frequency upper limit considered for the AR spectra
% g is a text string which represents a word on the word list considered
% nS is the number of speakers tested
N = 300;
ff = 2400;
nS = 21;
xax = (0:ff/(N-1):ff);

% Cut the AR spectra from 512 to N length
CC = AA(1:N,:); clear AA
BB = BB(1:N,:);

% Set up test and reference groups
%%    First sixteen reference group
AA=[CC(:,1:16), CC(:,32:47)];

% Compute the reference model
REF = mean(AA');
%%    Test group associated to reference group selected and the reference model
TG = [CC(:,17:31), BB(:,1:6), CC(:,48:62), BB(:,7:12), REF'];

% Run performanc measures
Itk
CB
Bnd
ccdist
    urte = ur;
    rrte = rr;
    TG = AA;
ccdist
    ure = ur;
    rre = rr;
```

```
% Check Code
%%    Itakura reference should equal zero
IdSdSr = fliplr(dSdSr);
Itakura_ref = IdSdSr(1);

%%    Cross-correlation coefficients references should equal one
CCrrf = fliplr(rrte);
CCurf = fliplr(urte);
CrosRef = [CCrrf(1) CCurf(1)];

%%    City block metric reference should equal 0
DifRef = fliplr(d);
Diff_reff = DifRef(1);

% Set-up check
CHECK = round([Itakura_ref Diff_reff CrosRef(1) CrosRef(2)]);
CHECKcheck = [0 0 1 1];

if CHECK == CHECKcheck
    [g, ' CHECKS GOOD']
else
    [g, ' CHECKS BAD']
end

% Display Bound Results
Fail_Bnd

% Display Itkura distance results
Imax = max(dSrdSr);
Fail_Ik = find(dSdSr>Imax)

% Display cross-correlation coefficients results
CrCor_ur = find(urte<min(abs(ure)))
CrCor_rr = find(rrte<min(abs(rre)))

% Display city block metric results
CB = find(d>max(de))
```

LIST OF REFERENCES

[1] Flege, J. E., and Hillenbrand, J., "Limits on phonetic accuracy in foreign language speech production," *The Journal of the Acoustic Society of America*, v. 76, no. 3, pp. 708-721, September 1984.

[2] Flege, J. E., and Eefting, W., "Cross-language switching in stop consonant perception and production by Dutch speakers of English," *Speech Communication*, v. 6, pp. 185-202, September 1987.

[3] Deller, J. R., Proakis, J. G., and Hansen, J. H. L., *Discrete-Time Processing of Speech Signals*, New York, New York, Macmillan Publishing Company, 1993.

[4] Carrell, J., and Tiffany W., *Phonetics: Theory and Application to Speech Improvement*, New York, New York, McGraw-Hill, 1960.

[5] Peterson, G., and Barney, H., "Control methods used in a study of vowels," *The Journal of the Acoustic Society of America*, v. 24, no. 2, pp. 175-184, March 1952.

[6] Therrien, C. W., *Discrete Random Signals and Statistical Signal Processing*, Englewood Cliffs, New Jersey, Prentice Hall, 1992.

[7] Strum, R. D., and Kirk, D. E. , *First Principles of Discrete Systems and Digital Signal Processing*, Reading, Massachusetts, Addison-Wesley Publishing Company Inc., 1988.

[8] VanDerKamp, M. M., *Modeling and Classification of Biological Signals*, MSEE Thesis, Naval Postgraduate School, Monterey, California, December 1992.

[9] Fargues, M. P., and Hippenstiel R., *Investigation of Spectral Based Techniques for Classification of Wideband Transient Signals*, Naval Postgraduate School Technical Report NPSEC-93-008, pp. 44-50, March 1993.

[10] Rabiner, L., and Juang, B-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, New Jersey, Prentice-Hall, 1993.

[11] Brown, D. W., and Fargues, M. P., *An Interactive Matlab Package for Signal Modeling and Analysis and Communications (with Speech Analysis and Linear Systems Modeling)*, Naval Postgraduate School Technical Report NPSEC-93-017, October 1993.

BIBLIOGRAPHY

1. Brown, D. W., and Fargues, M. P., *An Interactive Matlab Package for Signal Modeling and Analysis and Communications (with Speech Analysis and Linear Systems Modeling)*, Naval Postgraduate School Technical Report NPSEC-93-017, October 1993.

2. Carrell, J., and Tiffany W., *Phonetics: Theory and Application to Speech Improvement*, New York, New York, McGraw-Hill, 1960.

3. Deller, J. R., Proakis, J. G., and Hansen, J. H. L., *Discrete-Time Processing of Speech Signals*, New York, New York, Macmillan Publishing Company, 1993.

4. Fargues, M. P., and Hippenstiel R., *Investigation of Spectral Based Techniques for Classification of Wideband Transient Signals*, Naval Postgraduate School Technical Report NPSEC-93-008, pp. 44-50, March 1993.

5. Fargues, M. P., *Speech Processing*, ECE 4410 Class Notes, U. S. Naval Postgraduate School, Monterey, California, Spring 1993.

6. Flege, J. E., "The detection of French accent by American listeners, " *The Journal of the Acoustical Society of America*, v. 76, no. 3, pp. 692-707, September 1984.

7. Flege, J. E., and Eefting, W., "Cross-language switching in stop consonant perception and production by Dutch speakers of English," *Speech Communication*, v. 6, pp.185-202, September 1987.

8. Flege, J. E., and Hillenbrand, J., "Limits on phonetic accuracy in foreign language speech production," *The Journal of the Acoustic Society of America*, v. 76, no. 3, pp. 708-721, September 1984.

9. Fletcher, H., *Speech and Hearing in Communication*, New York, New York, D. Van Nostrand Company, Inc., 1953

10. Peterson, G., and Barney, H., "Control methods used in a study of vowels," *The Journal of the Acoustic Society of America*, v. 24, no.2 , pp. 175-184, 1952.

11. Rabiner, L., and Juang, B-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, New Jersey, Prentice-Hall, 1993.

12. Rabiner, L. R., and Schafer, R. W., *Digital Processing of Speech Signals*, Englewood Cliffs, New Jersey, Prentice-Hall, 1978.

13. Strum, R. D., and Kirk, D. E. , *First Principles of Discrete Systems and Digital Signal Processing*, Reading, Massachusetts, Addison-Wesley Publishing Company Inc., 1988.

14. Therrien, C. W., *Discrete Random Signals and Statistical Signal Processing*, Englewood Cliffs, New Jersey, Prentice Hall, 1992.

15. VanDerKamp, M. M., *Modeling and Classification of Biological Signals*, MSEE Thesis, Naval Postgraduate School, Monterey, California, December 1992.

16. Zue,V. W., "The Use of Speech Knowledge in Automatic Speech Recognition," *Proceedings of the IEEE*, v. 73, no. 11, pp. 1601-1615, November 1985.

INITIAL DISTRIBUTION LIST

	No. Copies
1. Defense Technical Information Center Cameron Station Alexandria, Virginia 22304-6145	2
2. Library, Code 52 Naval Postgraduate School Monterey, California 93943-5101	2
3. Chairman, Code EC Department of Electrical and Computer Engineering Naval Postgraduate School 833 Dyer Road, Room 437 Monterey, California 93943-5121	1
4. Professor Monique P. Fargues, Code EC/Fa Department of Electrical and Computer Engineering Naval Postgraduate School 833 Dyer Road, Room 437 Monterey, California 93943-5121	3
5. Professor Ralph Hippenstiel, Code EC/Hi Department of Electrical and Computer Engineering Naval Postgraduate School 833 Dyer Road, Room 437 Monterey, California 93943-5121	1
6. CPT John K. Dewey USA ELM JT ELTRWFAR (W4DJAA) Kelly, Texas 78241	2